DOCUMENT RESUME

ED 382 646 TM 023 080

AUTHOR Eignor, Daniel R.; And Others
TITLE Case Studies in Computer Adaptive Test Design through Simulation.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-93-56
PUB DATE Nov 93
NOTE 74p.; Version of a paper presented at the Annual Meeting of the National Council on Measurement in Education (Atlanta, GA, April 13-15, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Adaptive Testing; Algorithms; Case Studies; *College Entrance Examinations; *Computer Assisted Testing; Computer Simulation; Higher Education; Item Banks; *Item Response Theory; Selection; *Test Construction; Test Items; Test Reliability
IDENTIFIERS Calibration; *Graduate Record Examinations; Paper and Pencil Tests; Scholastic Aptitude Test; Three Parameter Model

ABSTRACT

The extensive computer simulation work done in developing the computer adaptive versions of the Graduate Record Examinations (GRE) Board General Test and the College Board Admissions Testing Program (ATP) Scholastic Aptitude Test (SAT) is described in this report. Both the GRE General and SAT computer adaptive tests (CATs), which are fixed length in nature, were developed from pools of items that were calibrated using the three-parameter item response theory model, and item selection was based on the recently developed weighted deviations algorithm (see Swanson and Stocking, 1992), which simultaneously deals with content, statistical, and other constraints in the item selection process. For the GRE General CATs (Verbal, Quantitative, and Analytical), item exposure was controlled by using an extension of an approach originally developed by Sympson and Hetter (1988). For the SAT CATs (Verbal and Mathematical), item exposure was controlled by using a less complex randomization approach. Lengths of the CATs were determined so that CAT reliabilities matched or exceeded comparable full length paper-and-pencil test reliabilities. Eight figures and 23 tables illustrate the analysis. (Contains 21 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-93-56

ED 382 646

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
R. Colgy
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

CASE STUDIES IN COMPUTER ADAPTIVE TEST DESIGN THROUGH SIMULATION

Daniel R. Eignor
Martha L. Stocking
Walter D. Way
Manfred Steffen

Educational Testing Service
Princeton, New Jersey
November 1993

TM020306

BEST COPY AVAILABLE

Case Studies in Computer Adaptive Test Design Through Simulation[1,2]

Daniel R. Eignor

Martha L. Stocking

Walter D. Way

Manfred Steffen

Educational Testing Service

---

[1]A previous version of this paper was presented at the annual meeting of NCME, Atlanta, 1993, at a Symposium entitled Practical Problems in the Development of Large Scale Computer Adaptive Tests.

[2]The SAT work described in this paper was supported by the College Board through Joint Staff Research and Development Committee funding and by the College Board and Educational Testing Service through Joint Planning Committee funding. The GRE work was supported by GRE Program funds.

Copyright © 1993. Educational Testing Service. All rights reserved.

Acknowledgements

Completion of the SAT and GRE General CAT development efforts described in this paper would not have occurred without the contributions of a large and varied number of staff. In the case of the SAT CAT, Gary Marco, Ted Blew, and Nancy Wright contributed to much of the initial planning and psychometric work and Ted, Nancy, and Rob Patrick were responsible for constructing the data base to support the CATs. John Dumont and Anne Connell were responsible for reviewing Verbal items for the initial pool to see if they were current and then they were responsible for building the SAT-V CAT item overlap lists. Al Schild did a similar review of the Math items to constitute the initial pool. Ed Curley and Gerry May provided the test development input that shaped the final characteristics of the SAT-V CAT. Jane Kupin and Jim Braswell provided similar input for the SAT-M CAT. Fred Schuppan ably assisted both Verbal and Math test development staff throughout the process and was responsible for production of paper-and-pencil copies of the CATs for review purposes.

In the case of the GRE General CAT, Clark Chalifour and Rob Patrick were responsible for constructing the data base to support the CATs. Jim Hessinger provided the test development input that shaped the final characteristics of the Verbal CAT; Fred Fischer provided similar input for the Quantitative CAT and Tim Habick for the Analytical CAT. Craig Mills, in his role as overall program director for the project, was responsible for much of the initial planning of work and for providing subsequent support.

Finally, review comments by Bill Ward and the efforts of Linda Ferner and Eugenia Tye in preparing the text, tables, and figures for this paper are greatly appreciated.

ABSTRACT

The extensive computer simulation work done in developing the computer adaptive versions of the Graduate Record Examinations (GRE) Board General Test and the College Board Admissions Testing Program (ATP) SAT is described in this report. Both the GRE General and SAT computer adaptive tests (CATs), which are fixed length in nature, were developed from pools of items that were calibrated using the three-parameter logistic IRT model and item selection was based on the recently developed weighted deviations algorithm (see Swanson and Stocking, 1992), which simultaneously deals with content, statistical, and other constraints in the item selection process. For the GRE General CATs (Verbal, Quantitative, and Analytical), item exposure was controlled by using an extension of an approach originally developed by Sympson and Hetter (1988). For the SAT CATs (Verbal and Mathematical), item exposure was controlled by using a less complex randomization approach. Lengths of the CATs were determined so that CAT reliabilities matched or exceeded comparable full length paper-and-pencil test reliabilities.

Case Studies in Computer Adaptive Test Design Through Simulation

Daniel R. Eignor

Martha L. Stocking

Walter D. Way

Manfred Steffen

INTRODUCTION

The evolution of theoretical and technological tools have made the implementation of computer adaptive testing (CAT) for large scale testing programs a reality. These large scale testing programs can be separated into two categories: those testing programs for which the CAT represents the only test to be implemented and administered, and those testing programs that already have full scale paper-and-pencil testing and now want to implement computer adaptive testing. The state of affairs is considerably more complicated for the latter set of testing programs, two of which are the subjects of the CAT development work to be described in this paper.

The College Board Admissions Testing Program (ATP) SAT and the Graduate Record Examinations (GRE) Board General Test are two examinations with rich paper-and-pencil testing traditions. Both Boards have recently decided to implement computer adaptive testing, although the manner of actual implementation differs considerably across the two Programs. (This will be discussed later in the paper.) In both cases, however, paper-and-pencil testing will continue and exist in tandem for some period of time with the computer adaptive testing. For the GRE General, duality of modes will exist until the paper-and-pencil test is eventually phased out. For the SAT, testing in both modes will continue until Spring 1994, when a new paper-and-pencil version of the SAT will be introduced. In either case, having two modes of testing necessitates that the CATs be developed using test specifications that are, to the extent possible, the same as those used to develop forms of the paper-and-pencil exams so that both exams are similar in terms of content tested and appearance to the examinee. Further, it necessitates that comparability of scores derived from the CAT and paper-and-pencil exams be established, so scores from both modes of testing can be used interchangeably (see Eignor, 1993).

The content of this paper deals with the issue of designing computer adaptive tests that are similar in content to their paper-and-pencil counterparts while, at the same time, having the adaptive tests achieve prespecified levels of important psychometric properties, such as acceptable levels of reliability. In addition, it may be necessary for the design of the CAT to incorporate other important features such as control of item exposure rates, procedures for administering items with common stimuli, and procedures for controlling overlap among items. This has been the case for both the SAT CAT and the GRE General CAT.

An excellent way of studying the variables just mentioned and other variables in the design of a CAT, and the interactions among the variables, is through the process of computer simulation. Certain design variables can be held fixed while others are left free to vary in such simulations. Results can be evaluated in terms of how well the resulting simulated CATs meet CAT content specifications and how well the simulated CATs do with respect to psychometric characteristics. Design variables, such as CAT test length, can be varied until the simulated CATs meet CAT content specifications and provide expected levels of reliability. In short, computer simulation provides the necessary mechanism for finalizing a number of decisions concerning CAT design before the CAT is actually implemented as part of an operational testing program.

The purpose of this paper is to describe the application of a CAT simulation system in the design of computer adaptive versions of the SAT and the GRE General Tests. This simulation system makes use of a recently developed procedure for selecting items in the computer adaptive testing environment, namely a weighted deviations model (see Swanson and Stocking, 1992). The next section of the paper provides background information on this model and algorithm, along with other necessary background information on SAT and GRE General paper-and-pencil test construction practices, i.e., the use of test specifications that must be paralleled in CAT test construction for these tests. Background information on procedures for controlling item exposure in CATs is also provided. Different procedures for controlling item exposure were used with the SAT and GRE General CATs. Finally, background information on the testing programs, the purposes of the CATs, and the content of the paper-and-pencil examinations is provided.

In the material that follows, reference will frequently be made to the SAT CAT. In reality, the SAT CAT is actually two separate CATs, one for SAT-Verbal and the other for SAT-Mathematical. In a similar fashion, the GRE General CAT is actually three separate CATs, one for GRE Verbal, one for GRE Quantitative, and one for GRE Analytical. Also, in the material that follows, reference will frequently be made to paper-and-pencil forms or tests, as for instance, a paper-and-pencil form of SAT-Verbal or the SAT-Verbal test. In reality, two sections of one overall six section SAT form constitute what is being referred to as an SAT-Verbal form or the SAT-Verbal test.

BACKGROUND[1]

SAT and GRE General Paper-and Pencil Test Construction Practices

SAT and GRE General Test paper-and-pencil test form construction is guided by a set of formal rules that govern whether or not an item may be included in the form being built. These rules are usually collectively referred to as "test specifications," and the rules constitute a set of constraints on the selection of items.

These constraints can be considered as falling into four separate categories: 1) constraints that focus on some intrinsic property of an item, 2) constraints that focus on item features in relation to all other candidate items, 3) constraints that focus on item features in relation to a subset of all other candidate items, and 4) constraints on the statistical properties of items as derived from pretesting.

Constraints on intrinsic item properties

Both the SAT and the GRE General Tests have explicit constraints on item content. For example, the test specifications for the Mathematics section of the SAT (SAT-M) specify the number or percentage of items on arithmetic, algebra, and geometry. These specifications are further elaborated by a specification that a certain percentage of the arithmetic items involve operations with whole numbers, a certain percentage involve fractions, and a certain percentage involve decimals. It is not unusual for fairly extensive test specifications to identify numerous content categories and subcategories of items and their required percentages or numbers; this is the case for both the SAT and the GRE General Test.

In addition to constraints explicitly addressing item content, constraints are typically given for other features intrinsic to an item that are not directly content related. For example, restrictions have been placed on the percentage of SAT-Verbal (SAT-V) sentence completion items that contain one blank as opposed to two blanks. These types of constraints treat the item type or the appearance of the item to the examinee. A second type of constraint not directly related to content may address the reference of the item to certain groups in the population at large, as when, for example, an SAT-V item with a science content has an incidental reference to a minority or female scientist. Such constraints may also seek to minimize or remove the use of items that contain incidental references that might appear to favor social class or wealth, for example, items dealing with country clubs, golf, polo, etc. These types of constraints are frequently referred to as "sensitivity" constraints and SAT and GRE General test specifications are designed to provide either a balance of such references or exclusion of such references in the interest of test fairness.

---

[1]Certain of the material contained in this section draws heavily on material appearing in a paper by Stocking and Swanson (1992).

Constraints that focus on item features in relation to all other candidate items

It seems obvious that SAT or GRE General test forms must not include an item that gives away the answer to another item. However, in addition to giving direct information about the correct answer to another item, an item can overlap with other items in more subtle ways. Items may test the same or nearly the same point but appear to be different at a casual glance, as a GRE General Quantitative item dealing with the sine of 90 degrees and the sine of 450 degrees. If the point being tested is sufficiently similar, then one item is redundant and is not included in the test because it provides no additional information about an examinee.

Items may also overlap with each other in features that are incidental to the purpose of the item. For example, two SAT-V reading comprehension passages might both be about science and both may contain incidental references to male scientists. It is unlikely that SAT-V test specialists would seek to include both passages on the test. Items that give away answers to other items, items that test the same point as others, and items that have similar incidental features are frequently referred to as exhibiting "content overlap", and such overlap must be constrained by the test specifications.

SAT-V or GRE General Verbal test specialists, when constructing verbal tests or test sections involving discrete verbal items, i.e., items that are not associated with a reading passage, are concerned that test specifications control a second kind of overlap, frequently referred to as "word overlap." The concern is that relatively uncommon words used in the stem or any of the answer choices for an item should not appear more than once in a test or test section. To do so would be to doubly disadvantage those examinees with more limited vocabularies in a manner that is extraneous to the purposes of the test.

Constraints that focus on item features in relation to a subset of all other candidate items

Some items are related to each other through their relationship to common stimulus material. This occurs when a number of items are based on a common reading passage in a SAT-V or GRE General Verbal section, or when a number of items are based on a common graph or table or figure in a GRE General Quantitative section. If test specifications dictate the inclusion of the common stimulus material, then some set of items associated with that material is also included in the test. It may be that there are more items available in a set than need to be included, in which case the test specifications dictate that some subset of the available items be included that best satisfy other constraints or test specifications. These groups of items are frequently referred to as "item sets" with the intended implication that items belonging to a set may not be intermixed with other items not belonging to the same set.

Constraints on the statistical properties of items

Information about the statistical behavior of SAT and GRE General items is available from the pretesting of these items in the variable sections of these tests. Test specifications typically constrain the selection of items based on their statistical behavior in order to construct test forms that have desired measurement properties, and this is the case for both the SAT and the GRE General Test.

These constraints typically take the form of specifying some target aggregation of statistical properties, where the statistical properties may be based on conventional difficulty and discrimination indices or the counterpart characteristics of items found in IRT. If IRT item characteristics are employed, the target might be some combination of item characteristics, as for example, target test information functions. If conventional item statistics are used, which is the case, for instance, for the SAT, the target aggregation is specified in terms of a frequency distribution of conventional item difficulties (actually transformations of conventional item difficulties to the "delta" scale; see Henrysson, 1971) and a target mean discrimination level.

The Stocking/Swanson Weighted Deviations Model

The foundation used at Educational Testing Service (ETS) for incorporating extensive and complex test specifications or constraints into the construction of adaptive tests involves the application of a weighted deviations model for automated item selection (AIS). The weighted deviations model is described in detail in Swanson and Stocking (1992). This model was developed in the context of a number of conventional test assembly paradigms that have been proposed in the literature over the last ten years. Typically, these paradigms employ a combination of IRT, modern computers, and linear programming models. Exemplars of such paradigms can be found in work by Theunissen (1985), van der Linden (1987), van der Linden and Boekkooi-Timmiga (1989), and Ackerman (1989). The weighted deviations algorithm differs, however, in one important aspect--its underlying philosophy--from the other paradigms just mentioned.

The underlying philosophy of the weighted deviations model, which makes it ideally suited for the construction of CATs subject to a large number of constraints, is as follows: Test assembly is less concerned with optimizing some function of the items selected (for example, maximizing test information or minimizing test length) or even meeting all of the constraints of interest (the other procedures attempt to do one or both of these things), than it is with coming "as close as possible" to meeting all constraints simultaneously. Thus constraints, including statistical constraints, are thought of as more like "desired properties" than as true constraints. This approach recognizes the possibility of constructing a test that may lack all of the desired properties at the expected levels, but emphasizes the minimization of aggregate failures. Moreover, the model provides for the possibility that not all constraints are equally important to the test developer by incorporating explicit relative weights as part of the modeling of constraints. If the item pool is rich enough in items with

intrinsic item features of interest, then the resultant test selected by the weighted deviations algorithm will have all the desired properties.

With this model, the constraints are formulated as bounds on the number of items having specified properties. The constraints need not, and in general will not, divide the item pool into mutually exclusive subsets. Rather, each item can have many different features that satisfy many different constraints. Further, statistical constraints on item selection are treated just like any other constraints. The algorithm seeks to minimize the weighted sum of positive deviations from these constraints. It employs a successive item selection procedure that makes it especially appropriate to a paradigm such as adaptive testing.

The four types of constraints used in the construction of SAT and GRE General paper-and-pencil forms are implemented in the construction of the SAT and GRE CATs by the weighted deviations algorithm in the following ways.

Constraints on intrinsic item properties

The control of intrinsic item features is accomplished through the use of explicit constraints, that is, lower and upper bounds (which may be equal) on the desired number of items which possess a feature. If items have been coded to a sufficient level of detail, it is possible to control the second type of constraint on item selection, undesirable overlap among items, by the same mechanism. For example, items that test the same point, or items that have similar incidental features, can be assigned a common code and then a constraint specified that only one such item may be included in an adaptive test. Likewise, an item that gives away the answer to another item can be assigned the same code as the other item, and a constraint can then be imposed so that only one of those items is administered to an individual.

Constraints that focus on item features in relation to all other candidate items

In practice, it is likely to be extremely difficult to develop and implement an item coding scheme with sufficient level of detail so that all overlap can be controlled by the imposition of explicit constraints alone. Instead, another mechanism is usually employed--that of overlap groups. An overlap group consists of a list of items that may not appear together in the same adaptive test. Overlap groups do not have to imply transitivity of overlap. That is, item A may overlap with item B, and item B may overlap with item C, but that does not imply that item A overlaps with item C since the reasons for the overlap between A and B and the overlap between B and C may be different. An extension of this concept is that overlap groups do not imply mutually exclusive groups of items since, again, the items may overlap for different reasons. These overlap groups can be formulated fairly simply by clerical methods. The detection of word (as opposed to content) overlap is made relatively simple by employing computerized tools that use (fallible) morphological algorithms to identify overlapping words (TD/DC 5.0 User's Manual, 1991). The detection

of content overlap is more complex, but will be feasible in the near future with computerized tools employing thesaurus-based algorithms to identify content similarities.

Once formed, these groups are used by the item selection algorithm to avoid the selection of any item that appears in a group with an item already administered. This provides a simple and completely effective solution to the problem of avoiding overlapping items.

Constraints that focus on item features in relation to a subset of all other candidate items

Theunissen (1986, p. 387) suggested that sets of items based on a common stimulus could be incorporated into a maximum information adaptive testing paradigm by the use of a set information function as the sum of the item information functions for the items comprising that set. This approach is effective in the context of constructing tests made up entirely of sets of items based on a common stimulus where the items associated with a particular stimulus are fixed in advance of test assembly and where the number of items in each set is equal or approximately equal.

This approach must be modified for a test composed of a mixture of item sets and discrete items or for a test where the number of items in a set varies greatly across sets and when the items to be administered from the set of items associated with a common stimulus are not specified in advance.

The approach taken in the construction of the SAT and GRE General CATs is consistent with Theunissen's suggestion in that partial sums of item information functions are computed as items (including items from a set) are administered. This approach is useful for the incorporation of items sets whether based on common stimulus material or common directions or some other feature which requires that the administration of items belonging to a set not be interrupted by the administration of other items not belonging to the same set. Each item set is assigned a conceptual partition of the item pool (a block); items not belonging to sets are not considered to be in such a partition. Some blocks may be designated as reenterable with a fixed number of items to be administered at each entry. For example, with a block of 85 SAT-V analogy items, the requirement might be that three analogy items must be administered together in a test that was constrained to have six analogy items in all. (This was not the procedure actually followed with SAT-V analogy items, where the blocking was such that all six analogy items were administered together.) Other blocks may be designated as not reenterable with a fixed number of items to be administered, as in a set of 6 SAT-V reading comprehension items associated with a reading passage from which three items are to be administered. (This procedure was followed with the SAT-V CAT.)

Blocks are entered (or possibly reentered) by the selection of an item in that block 1) that contributes the most to the satisfaction of all other constraints, and 2) that does not appear in an overlap group containing an item already administered. Once within a block, items continue to be selected adaptively for administration based on their contribution to the satisfaction of all constraints and overlap, until the number of items to be administered at that entry into the block is reached. If the block is not reenterable, it is then removed from further consideration in the pool; if it is reenterable, then the block remains available.

Constraints on the statistical properties of items

The main psychometric feature of adaptive testing involves the selection of items that have optimum statistical properties for measuring a particular examinee's ability. In the context of the construction of the SAT and GRE General CATs, the lower and upper bounds for this constraint were set equal to some large positive number. When considering the statistical properties of items, the weighted deviations algorithm will select those items that have the largest item information function at the current estimate of the examinee's ability.

Controlling Item Exposure

Any scheme that seeks to control the exposure of items in computer adaptive testing employs mechanisms that override the item selection procedure in use, thus degrading the quality of the adaptive test. Longer tests are therefore required to achieve the level of efficiency obtained when only the item selection procedure in use governs the choice of the next item, but longer tests may be viewed as a reasonable exchange for greater item and test security. One frequently used scheme for controlling item exposure involves a randomization approach. Sympson and Hetter (1985), however, have recently developed a more systematic approach to controlling item exposure.

Randomization Approaches

A typical randomization approach is to select the first item randomly from a group of five, the second randomly from a group of four, the third randomly from a group of three, and the fourth randomly from a group of two. The fifth and subsequent items are chosen to be optimal (see McBride and Martin, 1983). The assumption underlying this approach is that, after some number of initial items, examinees will be sufficiently differentiated so that subsequent items will vary a great deal. Thus it is sufficient to control the exposure of early items while not controlling the exposure of later items.

Many variations on this theme are possible, of course, including the possibility of never choosing the next item optimally with certainty, that is, the minimum group size is always two or greater. This latter approach recognizes that in spite of randomization on initial items, examinees with similar abilities may receive many of the same items subsequently unless attempts are made to control the exposure of items administered later in the test.

The Sympson and Hetter Approach

The simple randomization procedure described above attempts to increase item security by indirectly reducing item exposure. Sympson and Hetter (1985) tackle the issue of controlling item exposure directly in a probabilistic fashion.

The procedure distinguishes between the probability P(S) that an item is selected as optimal in an adaptive test for an examinee randomly sampled from a typical group of examinees, and P(A|S), the probability that an item is administered, given that it has been selected. If an item is administered every time it is selected as the optimal item, the item might become overexposed. The procedure seeks to control the overall probability that an item is administered, P(A) = P(A|S)*P(S), and to insure that the maximum value over all P(A)s is less than some value r. This value r is the expected (not observed) maximum rate of item usage.

The conditional probability P(A|S) = k is some fraction that indicates the proportion of the time an item is selected that it should actually be administered. The exposure control parameters, k, one for each item, are determined through a series of simulations (described in Stocking, 1993) using an already established adaptive test design and simulees drawn from a typical distribution of ability.

Once the exposure control parameters have been established, they are used in the adaptive test as follows:

1) Select the next item for administration.

2) Generate a random number uniformly distributed between 0 and 1.

3) If the random number is less than or equal to the exposure control parameter for the selected item, administer the item.

4) If the random number is greater than the exposure control parameter for the selected item, do not administer the item, and remove it from the pool of remaining items for this examinee. Repeat this procedure for the next-most-optimal item. Continue until an item is found that can be administered.

The Sympson/Hetter methodology, as originally developed, seeks to control exposure at the item level only. The SAT and GRE item pools both contain sets of items based on some common stimulus material, as in items based on the same reading comprehension passage. Stocking (1993) extended the Sympson/Hetter procedure to develop exposure control parameters for both the stimulus material and the items. This approach, which follows the same logic as the original Sympson/Hetter procedure, is referred to as the Extended Sympson/Hetter (ESH) procedure.

Further Details on the Adaptive Testing Procedure and Other Psychometric Features

The psychometric basis of the adaptive testing algorithm used with the SAT and GRE General CATs is most similar to that of Lord (1977) in the sense that an item is considered to have optimum statistical properties if it is most informative at an examinee's current maximum-likelihood estimate of ability. The first item is chosen to be most appropriate for an ability level of about -1.0 on the ability metric. Maximum likelihood estimates of examinee ability, based on responses to all previous items, are used to select the most informative item for subsequent administration, subject to the constraints on content, overlap and item sets previously described. Procedures to control item exposure and improve item security are imposed throughout the process. For the SAT CATs, a count-down randomization scheme whereby the first item is randomly chosen from a list of the eight best items, the second item is randomly chosen from a list of the seven best items, and so forth, was imposed. With this scheme, the eighth and subsequent items are chosen to be optimal. For the GRE General CATs, a similar randomization scheme was originally tried in early simulations. However, the results of randomization were not satisfactory in that it was not possible to control item exposure to the extent desired. For this reason, the Extended Sympson/Hetter methodology was imposed, with the desired maximum rate of usage set at .2.

The final ability estimate, after the administration of a fixed number of items, is converted to an estimated true formula score (SAT) or estimated number right true score (GRE General) on a reference set of items using the test characteristic curve (Lord, 1980, equations 4.9 and 15.6). This reference set of items is actually an intact conventional paper-and-pencil form of the test referred to in this paper as the "reference test". For each CAT, the items on this form have been calibrated and placed on the same metric as the item pool. The existing raw-to-scale conversion for this reference form can be used to create an SAT or GRE scaled score.

Early in the development process, a decision was made that both the SAT CATs and the GRE General CATs would be fixed rather than variable length CATs. The rationale for this decision is based on earlier CAT simulation work performed by Stocking (1987). In this earlier work, Stocking found that a bias was introduced into certain final ability estimates when CATs were allowed to be variable in length.

In summary, through use of the weighted deviations model, a mechanism exists for selecting items in the construction of the SAT and GRE adaptive tests that mirrors as closely as possible the considerations that govern the assembly of full-length paper-and-pencil forms of these tests. With this model, the next item administered in a fixed length adaptive test is the item that simultaneously

1) is the most informative item at an examinee's estimated ability level, and

2) contributes the most to the satisfaction of all other constraints in addition to the constraint on item information.

At the same time, it is required that the item

3) does not appear in an overlap group containing an item already administered, and

4) is in the current block (if in a block), starts a new block, or is in no block.

If 1) - 4) can be adhered to in the selection of items, the procedure provides for the selection of items that best satisfy the weighted deviations algorithm. However, if item exposure controls, such as randomization or the Extended Sympson/Hetter methodology, are imposed to increase the security of the CAT items pools, the resulting item selection will be less than optimal and the CAT will need to be lengthened.

The Testing Programs and Purposes of the CATs

SAT

The Admissions Testing Program (ATP) Scholastic Aptitude Test (SAT) is administered seven times a year at regular national test centers, which are high schools that have been designated as testing sites. The test is used primarily for college admissions purposes. Typically, a completely new form of the SAT is given at each administration. Most used forms are disclosed, either to satisfy legislation or to provide students with practice material for taking the SAT.

The SAT is currently being reconfigured and a new version of the test will be made available in 1994. The Verbal section of the test will no longer contain the antonym item type and the proportion of reading material is being increased. In addition, the lengths of the reading passages on the test will be increased to make them more like material normally read by high school students. The Mathematics section of the test will contain a new item type, the grid-in item, and calculators will be allowed for the first time. In short, the SAT is now entering a period of change.

However, with all of the activity going on at ETS and elsewhere related to the computerization of tests, the College Board decided, even though changes were being made to the SAT, that it was important to develop a computer adaptive prototype of the current test. One important difference, however, between the SAT CAT prototype and other adaptive tests being developed at ETS, such as the GRE General CAT, is that the SAT CAT was never intended to be used operationally, i.e., to yield scores to be used in the college admissions process. This was for two reasons:

1) The Program did not have a pool of secure items that could be devoted to the CAT. Hence, the CAT pool had to be built from items that had appeared on SAT paper-and-pencil forms that had been disclosed.

2) Even if a pool of secure items had existed for CAT purposes, no delivery mechanism was in place in the high schools to deliver the SAT CAT to the many students who would want to take it during the school year.

The SAT CAT prototype was developed with the intention that it would be gradually introduced into selected high schools. The purpose of the SAT CAT prototype is to provide students with a quick, yet novel, way to get an indication of how well they would do on the current full-length paper-and-pencil SAT. This can be contrasted with the GRE General CAT, which was developed from the start with the intention that scores be used in the graduate school admissions process.

GRE General

The paper-and-pencil version of the Graduate Record Examinations (GRE) General Test is administered five times a year at regular domestic test centers. The test is used primarily for graduate school admissions purposes. Typically, three of the five forms administered in a particular year are disclosed following their administration.

The GRE Board has been, and continues to be, interested in assessing a broader range of skills than currently are assessed on the paper-and-pencil GRE General Test. Recent technological developments involving computer administration have provided a mechanism to assess this broader range of skills, thereby hopefully enhancing the probability that the GRE General examination will identify persons likely to perform well in graduate school.

The move to computerization with the GRE General Test has taken place in stages. The first stage has involved the administration of secure linear computer-based GRE forms (i.e., paper-and-pencil forms administered via computer), to evaluate examinee acceptance and comfort with the computer delivery system. This first stage is viewed as temporary, with a move to the second stage, administration of adaptive tests that are similar in content to the paper-and-pencil forms of the test, to take place over the next one to two years. During this stage, paper-and-pencil and adaptive testing will exist simultaneously, with scores from both types of examinations used in the graduate school admissions process. In the third and ultimate stage, paper-and-pencil testing will be discontinued and all testing will be done on the computer via the adaptive process.

The movement from paper-and-pencil testing to CAT with the GRE General Test is occurring for several reasons. First, the amount of time currently devoted to a GRE General test administration is considered fixed. In order to expand the range of skills to be assessed, a goal of the GRE Board, it is necessary to reduce the amount of time allocated to the three measures that currently constitute the GRE General Test. CAT provides a logical way to maintain current psychometric characteristics for the three measures with fewer items and less testing time, thereby freeing up testing time to be devoted to the newer skills to be assessed. Second, as mentioned earlier, the majority of the measures to be used to assess the new skills under consideration for inclusion in the GRE General require or benefit from a

computer delivery system. Hence, the ultimate goal of the GRE Program is to have a set of adaptive tests, three of which will test the current three GRE General measures and some of which will test new skills, and for which there will be no paper-and-pencil counterparts.

SAT and GRE General Paper-and-Pencil Test Content

SAT

The current full-length paper-and-pencil SAT consists of six test sections, each with a thirty minute time limit. Two of the six sections contain SAT-Verbal items and two of the sections contain SAT-Mathematical items. One of the remaining sections contains the Test of Standard Written English (TSWE). The sixth section is a variable section that contains either pretest items or equating items.

One of the two SAT-Verbal (SAT-V) sections contains 45 items and the other contains 40 items. Scores are reported on the full 85 items. The 85-item test is made up of four item types, in the following numbers (in parentheses): reading comprehension items (25), antonym items (25), analogy items (20), and sentence completion items (15). The reading comprehension items are based on five or, more frequently, six passages. All SAT-Verbal items are five-choice multiple choice items. The raw score for each examinee on the 85-item SAT-V is created via formula scoring, using the formula R - ¼W for five-choice items. A conversion table exists for mapping the rounded raw (formula) score obtained on a particular form of SAT-V onto the 200 to 800 reporting scale.

One of the two SAT-Mathematical (SAT-M) sections contains 35 items and the other contains 25 items. Scores are reported on the full 60 items. The 60-item test is made up of two item types, in the following numbers (in parentheses): regular five-choice problem solving items (40) and four-choice quantitative comparison (QC) items (20). SAT-M is also broken down by content area. A 60-item SAT-M form contains items from the following four content areas, with numbers of items or ranges in parentheses: arithmetic items (18-19), algebra items (17), geometry items (16-17), and miscellaneous items (7-9).

The raw score for each examinee on the 60-item SAT-M is created via formula scoring, using the formula R -¼W for the five-choice items and R - ⅓W for the four-choice QC items. As with SAT-V, a conversion table exists for mapping the rounded raw (formula) score obtained on a particular form of SAT-M onto the 200 to 800 reporting scale.

Content specifications used in the development of forms of SAT-V and SAT-M involve much more elaborate breakdowns of items than either the item type breakdown (for SAT-V) or the item type and content breakdown (for SAT-M) just listed. For SAT-V, there are an additional 50 ways in which specifications on item types or other features are further broken down for test assembly purposes, while for SAT-M, there are a total of 114 additional ways in which the specifications on item type, content area, or other features are further broken down. These additional content specification breakdowns were used in the

development of SAT-V and SAT-M adaptive test constraints, which is discussed in a subsequent section of the paper.

GRE General

The current full-length paper-and-pencil GRE General test consists of seven test sections, each with a thirty minute time limit. Two of the seven sections contain GRE Verbal items, two contain Quantitative items, and two contain Analytical items. The seventh section is a variable section that contains pretest items or equating items.

Each of the two GRE Verbal sections are constructed to be parallel in appearance and each contains 38 items. Scores are reported on the full 76 items. The 76-item test is made up of four item types, in the following numbers (in parentheses): reading comprehension items (22), antonym items (22), analogy items (18), and sentence completion items (14). The reading comprehension items are based on four passages, with from 4 to 7 items per passage, depending on the length of the passage. All GRE Verbal items are five-choice multiple choice items. The raw score for each examinee on the 76-item GRE Verbal is a simple number-right score. A conversion table exists for mapping the number right score obtained on a particular form of GRE Verbal onto the 200 to 800 reporting scale.

Each of the two GRE Quantitative sections are constructed to be parallel in appearance and each contains 30 items. Scores are reported on the full 60 items. The 60-item test is made up of three item types, in the following numbers (in parentheses): regular five-choice problem solving items (20), four-choice quantitative comparison (QC) items (30), and five-choice data interpretation (DI) items (10). The data interpretation items appear in two sets, where each set contains 5 items. A 60-item Quantitative form contains items from the following three content areas, with ranges of items for each category in parentheses: arithmetic items (22-32), algebra items (13-21), and geometry items (12-20). Like GRE Verbal, the raw score for each examinee is a simple number-right score and a conversion table exists for each GRE Quantitative form to map number right scores onto the 200 to 800 reporting scale.

Each of the two GRE Analytical sections are constructed to be parallel in appearance and each contains 25 items[1]. Scores are reported on the full 50 items. The 50-item test is made up of two item types, in the following numbers (in parentheses): five-choice analytical reasoning items (38) and five-choice logical reasoning items (12). The analytical reasoning items appear in 6 sets, with each set containing from 3 to 8 items. As with the other GRE General Tests, the raw score for each examinee is a simple number-right score which is mapped onto the 200 to 800 reporting scale.

---

[1]Specifications for the GRE Analytical sections are currently being redefined. The specifications listed in this paper formed the basis for the development of the GRE Analytical CAT.

Content specifications used in the development of forms of the GRE General Test involve much more elaborate breakdowns of items than either the item type breakdowns (for Verbal and Analytical) or item type and content breakdowns (for Quantitative) just listed. For GRE Verbal, there are an additional 36 ways in which specifications on item types or other features are further broken down for test assembly purposes. For GRE Quantitative, there are 21 additional breakdowns and for GRE Analytical, there are 41 further subdivisions. These additional content specification and other breakdowns were used in the development of the GRE CATs.

METHOD

In each of the sections that follow, the SAT will be discussed first because the SAT CAT was developed before the GRE General CAT and, in a number of instances, work done on the SAT CAT directly informed comparable work done on the GRE CAT.

The Item Pools

SAT

Items from 18 disclosed SAT forms, or 1530 SAT-V items and 1080 SAT-M items, constituted the initial item pools for development of the SAT CAT. The 18 forms chosen had, for the most part, not received a major degree of exposure in the disclosure process, i.e., the forms had not been published in widely circulated books used to prepare students to take the SAT. Items on each of the 18 forms had been calibrated for IRT equating purposes using the 3-parameter logistic (3-PL) item response model and the computer program LOGIST (Wingersky, 1983). All items were calibrated on large samples (2000+) from current or recent SAT testing populations. The Stocking/Lord characteristic curve transformation method (see Stocking and Lord, 1983) was used to place item parameter estimates for SAT-V items from all 18 forms on the same scale. The same was done for the SAT-M items.

Each item in each of the initial pools was then screened using six criteria: 1) Was the content of the item still current?; 2) Did the item meet current ETS Sensitivity Guidelines?; 3) Did the item demonstrate obvious content overlap with another item or items in the pool?; 4) For items that had Differential Item Functioning (DIF) statistics, did the item exhibit an extreme level of DIF? (Some items came from forms administered prior to implementation of routine DIF analyses.); 5) Did the 3-PL IRT model exhibit poor fit to data for the item?; and 6) For SAT-V items only, did the item exhibit an extreme level of word overlap with other items in the pool? Items not meeting the six criteria were screened from the pools, leaving a total of 998 Verbal items and 863 Math items.

The SAT-V pool of 998 items and the SAT-M pool of 863 items were seen as unnecessarily large given that it was assumed the final CATs would be in the range of 20 to 30 items. Hence, each of the total pools was randomly split into initial active and inactive

half-pools of approximately 499 Verbal items and 431 Math items. In the ensuing series of sequential computer simulations (described in a subsequent section), items were borrowed from the inactive pools so that content or item type constraints that were not being met could be met with the active pools. Hence, the inactive SAT-V and SAT-M pools are, at this point, not suitable for CAT purposes.

After a number of sequential simulations, it was found that content, statistical, and other constraints for a 27-item SAT-V CAT and a 20-item SAT-M CAT could be met through the use of an SAT-V pool of 303 items and an SAT-M pool of 235 items. (Elaboration on the reasons for selecting CATs of 27 and 20 items will be presented in a subsequent section.) Table 1 contains the numbers of items in each of the SAT-V item type categories comprising the final 303 item Verbal pool. The 91 reading comprehension items are based on 27 passages, with each passage having from 3 to 6 items based on it. Also contained in Table 1 is the breakdown, by item type, of the number of items in the 27-item SAT-V CAT. The numbers of items for the item types on the CAT are basically proportional to the numbers of items for the item types that are contained on the full-length 85-item SAT-V paper-and-pencil test.

Insert Table 1 about here

Table 2 contains the numbers of items in the two SAT-M item type categories and the numbers of items in each of the four content categories comprising the final 235 item Math pool. Of the 235 items, there are 12 that appear in 6 sets of two items each; these items are all regular five-choice problem solving items. Each two item set shares a common stimulus. Also contained in Table 2 is a breakdown by item type and by content area of the number of items on the 20-item SAT-M CAT. As with the SAT-V CAT, these numbers are basically proportional to the numbers of items in the same categories on the full-length 60-item SAT-M paper-and-pencil test.

Insert Table 2 about here

For the final SAT-V pool, the mean estimated value of the item discrimination parameter ($\hat{a}$) was .95, with a standard deviation of .29, and a range from .30 to 1.83. The mean estimated item difficulty parameter ($\hat{b}$) was -.01, with a standard deviation of 1.29, and a range from -2.68 to 2.56. The mean estimated lower asymptote parameter ($\hat{c}$) was .16, with a standard deviation of .09, and a range of .01 to .50. For the final SAT-M pool, the mean value of $\hat{a}$ was 1.13, with a standard deviation of .32, and a range from .37 to 1.91. The mean $\hat{b}$ was .11, with a standard deviation of 1.18, and a range from -2.82 to 2.52. The mean $\hat{c}$ was .12, with a standard deviation of .09, and a range of 0 to .46. For both pools, an examination of the quality of the pool in terms of the information function for the entire pool indicated that the pool, in the aggregate, contained more information at ability levels above the average ability level of zero than below the average ability level.

GRE General

Unlike the SAT CAT pools, the three GRE General CAT item pools were built from secure items. Most of these items were newly written and pretested items, although some items from secure final forms were added to the pools. Prior to pretesting, each item was reviewed to make sure 1) that the content of the item was still considered current, and 2) the item met ETS Sensitivity Guidelines. Items were then included in the pretest or variable section of the GRE General Test and administered at regular domestic administrations; different pretest forms were spiralled in the variable section at these administrations.

Only pretest items that did not require any subsequent changes were eligible for the CAT pools. These items were calibrated using the 3-PL model and LOGIST; sample sizes were in excess of 2000 examinees. Items for which the 3-PL model exhibited poor fit to the data were not considered further. The Stocking/Lord characteristic curve transformation method was then used to put the pretest parameter estimates for the remaining items on a common scale.

After a reasonably large item pool for each of the three GRE General tests had been assembled, each item in each pool was reviewed again to see if the item demonstrated obvious content overlap with another item or items in the pool. After removing such items from each of the three pools, the remaining items constituted the pools taken into the initial simulations. The sizes of these pools were: Verbal - 526 items, Quantitative - 496 items, and Analytical - 578 items.

After a number of sequential simulations, it was found that content, statistical, and other constraints for a 30-item GRE Verbal CAT, a 28-item Quantitative CAT, and a 35-item Analytical CAT could be met through the use of a Verbal pool of 350 items, a Quantitative pool of 330 items, and an Analytical pool of 449 items. Table 3 contains the numbers of items in each of the Verbal item type categories comprising the final 350 item Verbal pool. The 185 reading comprehension items are based on 31 passages, with each passage having from 5 to 10 items based on it. Also contained in Table 3 is the breakdown by item type of the number of items on the 30-item GRE Verbal CAT. As with the SAT CATs, the numbers of items for the item types on the CAT are basically proportional to the number of items for the item types that are contained on the full-length 76-item GRE Verbal paper-and-pencil test.

Insert Table 3 about here

Table 4 contains the numbers of items in the three GRE Quantitative item type categories and the numbers of items in each of the three content categories comprising the final 330 item pool. The 129 data interpretation items in the pool are based on 18 sets, with each set having from 5 to 11 items. Also contained in Table 4 is a breakdown by item type

and by content area of the number of items in the 28-item Quantitative CAT. Again, these numbers are basically proportional to comparable numbers comprising the full length 60-item paper-and-pencil test.

Insert Table 4 about here

Table 5 contains the number of items in the two GRE Analytical item type categories comprising the final 449 item pool. The 374 analytical reasoning items are based on 61 sets, with each set having from 6 to 8 items. The 75 logical reasoning items sub-pool is made up of 69 discrete items and 3 sets of 2 items each. Also contained in Table 5 is the breakdown by item type of the numbers of items in the 35-item Analytical CAT. As with the other CATs, these numbers are basically proportional to comparable numbers comprising the full length 50-item paper-and-pencil test.

Insert Table 5 about here

For the final GRE Verbal pool, the mean estimated value of the item discrimination parameter ($\hat{a}$) was .80, with a standard deviation of .25, and a range of .28 to 1.58. The mean estimated value of the item difficulty parameter ($\hat{b}$) was -.46, with a standard deviation of 1.18, and a range of -3.73 to 2.25. The mean estimated value of the lower asymptote parameter ($\hat{c}$) was .16, with a standard deviation of .11, and a range from 0 to .50. For the final GRE Quantitative pool, the mean $\hat{a}$ was .91, with a standard deviation of .35, and a range of .26 to 1.84. The mean $\hat{b}$ was .02, with a standard deviation of 1.21, and a range of -4.64 to 2.46. The mean $\hat{c}$ was .13, with a standard deviation of .10, and a range of 0 to .50. For the final GRE Analytical pool, the mean $\hat{a}$ was .79, with a standard deviation of .25, and a range of .25 to 1.78. The mean $\hat{b}$ was -.04, with a standard deviation of 1.29, and a range of -4.49 to 4.40. The mean $\hat{c}$ was .16, with a standard deviation of .10, and a range of 0 to .50. Like the SAT pools, an examination of the quality of the Quantitative and Analytical pools in terms of the information function for the entire pool indicated that the pool, in the aggregate, contained more information at ability levels above the average ability level of zero than below the average ability level. For the GRE Verbal pool, the information function for the entire pool was just about centered on the average ability level of zero, and was symmetric for the most part around the average ability level so that comparable amounts of information were supplied above and below the average ability level.

Adaptive Test Constraints

SAT-V

Content Constraints

SAT-V items and passages to be selected for the CAT were classified by test development specialists according to 54 different features. These specialists specified the number of items desired for each feature, paralleling the process of assembling the current paper-and-pencil test. These 54 features formed the initial set of 54 content-related constraints for the SAT-V CAT. For reasons to be explained later, 13 of these constraints received zero weights, thus reducing the number of constraints to be actually dealt with to 41. These 41 constraints on item selection are listed in Table 6. The weighted deviations model actually employs a single constraint for every feature that has equal lower and upper bounds, and two constraints for every feature that has unequal lower and upper bounds. Thus, from the perspective of the weighted deviations algorithm, the specifications in Table 6 represent a total of 71 constraints [11 + (2 x 30)]. However, for ease of discussion, the test specialists' perspective of 41 constraints on item features will be adopted.

Insert Table 6 about here

Prior to the beginning of the series of SAT-V simulations, lower and upper bounds for all 54 constraints were specified for adaptive test lengths of 20, 21, . . ., 30 items, because it was hypothesized in advance that the final satisfactory test length would lie within this range. Shown in Table 6 are the lower and upper bounds for the constraints with non-zero weights for the final SAT-V adaptive test of 27 items. Next to these bounds are the relative weights assigned to the satisfaction of each constraint in the final test design; these weights reflect the relative importance of the constraint to the test specialists. In addition, the number of passages or items in the pool that are identified as having each specific property is listed.

The first 10 constraints listed in Table 6 are relevant to the length or content of reading passages. For example, two constraints are formed based on whether a passage is categorized as being long or medium in length--test specialists wanted two of the former and one of the latter to appear in each CAT. The next 5 constraints (Constraints 11-15) are relevant to the items associated with the reading passages. Constraints 16 through 24 are constraints on sentence completion items; constraints 25 through 33 are constraints relevant to analogy items; and constraints 34 through 41 are relevant to antonym items. Constraints 11,16, 25, and 34 specify the total number of items in each of these major types to be included in the CAT.

The constraint weights listed in Table 6 are those that were used in the final satisfactory test design of 27 items. The weight given the constraint on item information, computed at 21 different ability levels for each item, was 15. For the selection of a single item for a single examinee, however, only the ability level closest to the current proficiency estimate is considered. Thus, from the perspective of the weighted deviations algorithm, the statistical constraint adds only one additional constraint on item selection.

The weights in Table 6 were arrived at through an iterative trial-and-error process involving a series of sequential simulations, where constraint weights were specified, the resulting adaptive tests were examined for constraint violations, and some weights were changed to reduce important violations. Constraints with the highest weight, 20, are so important that they cannot be violated and the resultant adaptive test be judged acceptable. Others receive lower weights because, although they are considered to be important, some constraint violations may be acceptable.

As mentioned earlier, thirteen constraints had zero weights, thus removing these constraints from the CAT design problem. This was done for varying reasons. Certain of the constraints were associated with reading passages. The purpose of these constraints was to attempt to insure, for both medium and long reading passages, that examinees received items on information contained in the first half of the passage, the second half of the passage, and the passage as a whole. This is in contrast to the situation where, for example, all items associated with a passage ask about information contained only in the first half of the passage. These constraints were removed because no single reading passage had associated with it items of all possible types, thus constraint violation was inevitable. If these constraints are important to satisfy, the item pool must be augmented with many more passages with many more items of all types associated with them; this was not seen as feasible. Some constraints were removed from the problem because there were so few items in the pool that the constraint was almost never violated anyway, or because upon reconsideration by test specialists, the constraint became viewed as unimportant.

The 41 constraints with nonzero weights in Table 6, plus the constraint on information and the constraints on overlap and item sets to be discussed next, constitute the set of desired properties that the weighted deviations algorithm attempted to satisfy in the selection of items for the SAT-V adaptive test.

Overlap Constraints

Table 7 gives a portion of the set of overlap groups constructed by test specialists after careful examination of the SAT-V pool. Items may be indicated as overlapping with other items and/or with passages. Passages may be indicated as overlapping with other passages and/or discrete items. If a passage overlaps with another passage or with a discrete item, all of the items associated with the passage(s) are considered to overlap. The entries listed in each overlap group indicate items and passages that may not be administered

together in the same adaptive test. For this pool of 303 items and 27 passages, there was a total of 547 such groups with 1,926 entries.

Insert Table 7 about here

Item Set Constraints

Table 8 displays a portion of the list of blocks of items that are to be considered in sets. For SAT-V , none of the blocks are reenterable and every item appears in a block. Test specialists felt that to enhance comparability with the conventional paper-and-pencil test, it was necessary to administer all sentence completion items together; likewise all antonyms and all analogies. Reading comprehension passages and items can appear anywhere within the test (except at the very beginning), but once started, cannot be interrupted. For this pool, there is a total of 54 logical blocks.

Insert Table 8 about here

SAT-M

Content Constraints

SAT-M items and sets to be selected for the CAT were classified by test development specialists according to 120 different features. (SAT-M specialists attempted to deal with the issue of overlap among items through the use of a large number of features, i.e., content constraints, rather than through the development of overlap constraints.) These specialists specified the number of items desired for each feature, paralleling the process of assembling the current paper-and-pencil test. These 120 features formed the initial set of 120 constraints for the SAT-M CAT. For reasons to be explained later, 45 of these constraints received zero weights, thus reducing the number of constraints to be dealt with to 75. These 75 constraints on item selection are listed in Table 9. As mentioned for SAT-V,.these actually represent a total of 139 constraints [11 + (2 x 64)] to be evaluated by the weighted deviations algorithm, but, again, the test specialists' perspective of 75 constraints will be used.

Insert Table 9 about here

At the beginning of the series of SAT-M simulations, lower and upper bounds for all 120 constraints were specified for adaptive test lengths of 18, 19, . . ., 25 items because it

was hypothesized in advance that the final satisfactory test length would lie in this range. (For the SAT-V CAT simulations done first, the final adaptive test length of 27 items is approximately one-third the length of the paper-and-pencil test. Hence, it was hypothesized that the final SAT-M CAT length would be about a third of the 60 items on the paper-and-pencil test, or around 20 items.) Shown in Table 9 are the lower and upper bounds for the 75 constraints with non-zero weights for the final SAT-M adaptive test length of 20 items. Next to these bounds are the relative weights assigned to the satisfaction of each constraint in the final test design. In addition, the number of sets or items in the pool that are identified as having each specific property is listed.

Constraint 1 has to do with the item sets; it is specified that there is to be either zero or one item set on each CAT and there are six item sets (each containing two regular five-choice items) from which to choose. Constraints 2-9 pertain to items in all four content categories (arithmetic, algebra, geometry, and miscellaneous) and in both type categories (five-choice regular problem solving and four-choice quantitative comparison). Constraints 10-15 are relevant to geometry items in both type categories. Constraints 16-24 pertain to the four-choice quantitative comparison items. The remaining constraints deal with the five-choice regular problem solving items and specifically with these items for each of the four content categories. Constraint 25 deals with all five-choice items, while constraints 26-41 deal with five-choice arithmetic items, 42-52 with five-choice algebra items, 53-70 with five-choice geometry items, and finally, 71-75 with five-choice miscellaneous items. Constraints 16 and 25 specify the total number of items of each of the two item types to be included in the CAT. Constraints 17-20 specify the numbers of quantitative comparison items by content category to be included on the CAT. Constraints 26, 42, 53, and 71 specify the numbers of regular five-choice problem solving items by content category to be included on the CAT.

The constraint weights listed in Table 9 are those that were used in the final SAT-M CAT test design of 20 items. The weight given the constraint on item information, computed at 21 different ability levels for each item, was 9. As with SAT-V, this statistical constraint adds only one additional constraint on item selection.

As with SAT-V, the weights in Table 9 were arrived at through an iterative trial-and-error process involving a series of sequential simulations, where constraint weights were specified, the resulting adaptive tests were examined for constraint violations, and some weights were changed to reduce constraint violations. As with SAT-V, those constraints that were deemed the most important to be met were given the highest weight--20. Other constraints were given lower weights because some constraint violations were viewed as acceptable.

As mentioned earlier, 45 constraints were given weights of zero, thus removing them from the CAT design problem. Unlike SAT-V, this was done for a single reason. The SAT-M constraints, unlike the SAT-V constraints, involve a good number of constraints embedded within other constraints, i.e., one constraint is a special case of the other. So whereas one constraint might involve five-choice geometry items that deal with angles in a

plane, special cases or additional constraints embedded within this constraint might deal with right angles in a plane, complementary angles in a plane, etc. This fine gradation or layering of constraints was considered in the CAT design in an attempt, as mentioned before, to circumvent the development of overlap constraints. However, for many of these embedded constraints no appropriate items existed in the pool. Hence, these constraints were given weights of zero.

The 75 constraints in Table 9, plus the constraint on information and the constraints on item sets to be discussed next, constitute the set of desired properties that the weighted deviations algorithm attempted to satisfy in the selection of items for the SAT-M adaptive test.

Item Set Constraints

The list of blocks of items for SAT-M is much less detailed than that for SAT-V and will not be presented in a table. Essentially there is a block for each of the six five-choice item sets, a block for the 107 four-choice quantitative comparison items in the pool and a block for the 128 regular five-choice problem solving items in the pool. The regular five-choice problem solving block is reenterable, but the other blocks are not. Test specialists felt that it was important that all seven quantitative comparison items on the CAT be administered together, but that the thirteen regular five-choice problem solving items could be administered either as one long block or as two shorter blocks. Of course, the items for a particular set had to be given together because these items share a common stimulus.

GRE Verbal

Content Constraints

GRE Verbal items and passages to be selected for the CAT were classified by test development specialists according to 40 different features. The specialists specified the number of items desired for each feature, and the 40 features were changed into 40 different constraints on item or passage selection. Four constraints were subsequently removed (i.e., given zero weights) and the remaining 36 constraints are listed in Table 10. As with SAT-V and SAT-M, these actually represent a total of 62 constraints [10 + (2 x 26)] to be evaluated by the weighted deviations algorithm but, again, the test specialists' perspective of 36 constraints will be used.

Insert Table 10 about here

At the beginning of the series of GRE Verbal simulations, lower and upper bounds for the 40 constraints were specified for adaptive test lengths of 17, 18, 19, . . . , 37 items. Shown in Table 10 are the lower and upper bounds for the 36 constraints with non-zero

weights for the final GRE Verbal adaptive test length of 30 items. Next to these bounds are the relative weights assigned to the satisfaction of each constraint on the final test design. The number of items or passages in the pool that are identified as having each specific property is also listed.

Constraints 1 and 2 have to do with long and short reading comprehension passages; it is specified that there be one long reading passage and two short reading passages on the CAT. Constraints 3-9 deal with other features of the reading comprehension passages. Constraints 10-14 deal with features of the reading comprehension items; constraints 15-19 with features of the sentence completion items; constraints 20-24 with features of the analogy items; and, finally, constraints 25-29 with features of the antonym items. The remaining constraints, 30-36, deal with features common to all four item types. Constraints 10, 15, 20, and 25 specify the numbers of items from the four item type categories to appear on the GRE Verbal CATs.

As mentioned earlier, the constraint weights listed in Table 10 are those used in the final GRE Verbal CAT test design of 30 items. The weight given the constraint on item information, computed at 21 different ability levels for each item, was 20. This statistical constraint adds only one additional constraint on item selection.

The weights in Table 10 were arrived at through an iterative trial-and-error process in the same way the weights were arrived at for SAT-V and SAT-M. In the case of GRE Verbal, the constraint on item information was given the highest weight--20. Passage-related constraints were given the next highest weight--15. Other constraints were given lower weights because some constraint violations were viewed as acceptable.

The 36 constraints in Table 10, plus the constraint on information and the constraints on overlap and item sets to be discussed next, constitute the set of desired properties that the weighted deviations algorithm attempted to satisfy in the selection of items for the GRE Verbal adaptive test.

Overlap Constraints

Overlap constraints for the GRE Verbal CAT were set up like those for SAT-V, and an overlap table like that presented for SAT-V (see Table 7) was constructed. It, too, involved both items and passages, and the logic was the same as for SAT-V. For the GRE Verbal pool of 350 items and 31 passages, there was a total of 169 overlap groups and 382 entries.

Item Set Constraints

For GRE Verbal, only the reading comprehension items that could be administered with particular reading comprehension passages were blocked, and none of these blocks were reenterable. Since there are 31 passages, there were 31 blocks. Unlike SAT-V, no blocking

was done for GRE Verbal on the other item types. Hence, for example, the six GRE Verbal sentence completion items specified do not have to appear together as a group. This is one important way in which the GRE Verbal CAT differed from the SAT-V CAT. In the SAT-V CAT, the various item types occur together in blocks.

GRE Quantitative

Content Constraints

GRE Quantitative items and sets to be selected for the CAT were initially classified by test development specialists according to 22 different features. These specialists specified the number of items desired for each feature, paralleling the process of assembly of the current paper-and-pencil test. Two additional features were subsequently added, and the total of 24 features formed the set of 24 constraints on item selection for GRE Quantitative presented in Table 11. These actually represent a total of 37 constraints [11 + (13 x 2)] to be evaluated by the weighted deviations algorithm, but the test specialists' perspective of 24 constraints will be retained.

Insert Table 11 about here

At the beginning of the series of GRE Quantitative simulations, lower and upper bounds were specified for all constraints for adaptive test lengths of 10, 11, 12, . . . 30 items. Shown in Table 11 are the lower and upper bounds for the constraints for the final Quantitative adaptive test length of 28 items. Next to these bounds are the relative weights assigned to the satisfaction of each constraint in the final test design. Also listed is the number of sets or items in the pool that are identified as having each specific property.

Constraint 1 has to do with the data interpretation (DI) item sets; it is specified that there be two such sets on each CAT. Constraints 2-4 have to do with the quantitative comparison (QC) items in the three content categories, while constraints 5-7 have to do with the problem solving (PS) items in the same three content categories. Constraints 8-17 have to do with a particular item type (Type 1) and the numbers of Type 1 DI, QC, and PS items to appear on the CATs. Constraints 18-20 pertain to all items, while constraints 21-24 pertain to the QC items only. The sum of the items specified for constraints 2-4 give the total number of QC items to appear on the CAT. The sum of the items specified for constraints 5-7 give the total number of PS items to appear on the CAT. Finally, constraint 9 specifies the number of DI items to appear on the CAT.

The constraint weights listed in Table 11 are, as mentioned earlier, the weights used in the final GRE Quantitative test design of 28 items. The weight given the constraint on item information, computed at 21 different ability levels for each item, was 10. This statistical constraint adds only one additional constraint on item selection.

As with the other CATs, the weights in Table 11 were arrived at through an iterative trial-and-error process. In the case of the Quantitative CAT, the highest weight, 11, was placed on the constraint dealing with the data interpretation sets, although a number of other constraints received weights of 10. The remaining constraints were given lower weights because some violations of these constraints were viewed as acceptable.

The 24 constraints in Table 11, plus the constraint on information and the constraints on overlap and item sets to be discussed next, constitute the set of desired properties that the weighted deviations algorithm attempted to satisfy.

Overlap Constraints

Overlap constraints for GRE Quantitative were set up just like those for SAT-V or GRE Verbal, except that sets rather than passages were involved. For the Quantitative pool of 330 items and 18 sets, there was a total of 57 overlap groups and 198 entries.

Item Set Constraints

For GRE Quantitative, only the DI items that could be administered as part of a DI set were blocked, and none of these blocks were reenterable. Since there are 18 sets, there were 18 blocks. Unlike SAT-M, the QC items were not blocked and hence the QC items could be interspersed throughout the CAT with the PS items. For SAT-M, it was specified that the 7 QC items on that CAT appear together as a group or block.

GRE Analytical

Content Constraints

GRE Analytical items and sets to be selected for the CAT were initially classified by test development specialists according to 47 different features. The specialists specified the numbers of items desired for each feature, and the 47 features were changed into 47 different constraints on item or passage selection. Eight of the constraints were subsequently removed, either because there were not enough items in the pool to satisfy the constraint or the constraint was later seen as unimportant, and the remaining 39 constraints are listed in Table 12. These actually represent a total of 74 constraints [4 + (35 x 2)] to be evaluated by the weighted deviations algorithm, but as with the other CATs, the test specialists' perspective of 39 constraints will be retained.

Insert Table 12 about here

At the beginning of the series of GRE Analytical simulations, lower and upper bounds for all constraints were specified for adaptive test lengths of 31, 32, 33, . . . 40 items. Shown in Table 12 are the lower and upper bounds for the 39 constraints for the final GRE Analytical test length of 35 items. Next to these bounds are the relative weights assigned to the satisfaction of each constraint in the final test design. The number of items or sets in the pool that are identified as having each specific property is also listed.

Constraints 1-12 deal with characteristics of the Analytical Reasoning item sets. Constraint 13 specifies the number of analytical reasoning items to appear on a CAT and constraint 14 specifies the number of logical reasoning items to appear. Constraints 15-32 deal with features of the logical reasoning items. Constraints 33-39 deal with features involving all items (in this case, the two specific item types).

As mentioned earlier, the constraint weights listed in Table 12 are those used with the final GRE Analytical CAT test design of 35 items. The weight given the constraint on item information, computed at 21 different ability levels for each item, was 20. This statistical constraint adds only one additional constraint on item selection.

The weights in Table 12 were arrived at through the same iterative trial-and-error process used with the other CATs. The constraints that were deemed the most important to be met, having to do with the numbers of analytical reasoning sets, analytical reasoning items, and logical reasoning items, were given the highest weight--30. For reasons explained below, a number of the other Analytical constraints also received large weights.

An examination of information in selected tables highlights at least two distinctions between the GRE Analytical CAT and the other four CATs. Looking at information in Tables 8-12, it can be seen that the weights associated with the Analytical constraints are uniformly higher than the comparable weights for the other four CATs. Looking at information contained in Tables 1-5 and in the text, it can be seen that a larger proportion of items in the initial Analytical CAT pool were used in the final simulation run then was the case with the other CATs. One plausible explanation for this is that the Analytical CAT is the most set dependent of the five CATs described in this paper. In order to adequately control the operation of the weighted deviations algorithm, it was necessary to weight constraints associated with Analytical Reasoning sets relatively heavily. With these relatively large sets of weights, it became necessary to increase the weights associated with the Analytical Reasoning and Logical Reasoning item types so that those constraints could be met. Finally, while the Analytical CAT contains 35 items, in reality there are only 15 independent selections for administration (the 9 logical reasoning items and the 6 analytical reasoning sets). Since exposure rates are controlled with the Analytical CAT for sets as well as items, as the number of sets used naturally increases, so does the number of items used. This is why 449 of the initial 578 items in the Analytical pool were used in the final simulation run.

In sum, the 39 constraints in Table 12, plus the constraint on information and the constraints on overlap and item sets to be discussed next, constitute the set of desired properties the weighted deviations algorithm attempted to satisfy.

Overlap Constraints

Overlap constraints for GRE Analytical were set up in the same way as those for the other three CATs that had overlap lists. For the GRE Analytical pool of 449 items and 64 sets, there was a total of 106 overlap groups and 218 entries.

Item Set Constraints

For GRE Analytical, only the analytical reasoning items associated with the 61 analytical reasoning sets and the logical reasoning items associated with the 3 logical reasoning sets were blocked, and none of these blocks were reenterable. Hence, in total, there were 64 blocks.

Details of the Simulation Process Used with All CATs

The design phase for each of the five CATs described in this paper essentially followed the same set of iterative steps, although there were variations in details, such as procedures for controlling item exposure. Each CAT was designed to be as similar to existing paper-and-pencil forms as possible--both in terms of item type and content breakdowns and score precision--in the shortest possible fixed adaptive test length. In the case of the SAT-V and SAT-M, the goal was to achieve an estimated reliability equal to the average reliability of the ten most recent paper-and-pencil test forms. For the GRE CATs, both the conditional standard error of measurement and the estimated reliability were assessed in comparison to the paper-and-pencil reference test. The steps of the iterative process used with each CAT are as follows:

1. Initial weights for the content and other constraints and upper and lower bounds for the numbers of items for each constraint were set and an initial CAT test length was chosen.

2. For SAT-V and SAT-M, a count-down randomization scheme was put in place to control item exposure. With this scheme, the first item to be administered is randomly chosen from a list of the eight best items, the second item is randomly chosen from a list of the seven best items, and so forth. The eighth and subsequent items are chosen to be optimal. For the GRE General CATs, the Extended Sympson/Hetter methodology was imposed with the desired maximum rate of usage set at .2. (For initial simulations, a randomization scheme similar to that used for the SAT was also investigated.) It is worth mentioning that once the Extended Sympson/Hetter methodology was implemented for the GRE simulations, this step required a series of repeated simulations where the

exposure control parameters were adjusted in each replication, and the replications continued until the exposure control parameters stabilized.

3. For each test, an initial simulation was performed by replicating a fixed number of simulees at points equally spaced along the observed score metric. The number of equally spaced intervals varied according to the paper-and-pencil test lengths of each measure, ranging from 9 for GRE Analytical (where the observed number right score range extends from 0 to 50) to 19 for SAT-V (where the observed formula score range extends from below 0 to 85). The intervals typically covered abilities ranging from around chance to just below a perfect score. The number of replications ranged from 100 to 200, and varied partly because of varying demands on computer processing time in the simulations for the different tests. (Computer processing time especially became a factor in the GRE simulations that made use of the Extended Sympson/Hetter methodology.)

4. In the construction of the CATs in the initial simulation (and subsequent simulations), item set constraints and, in the case of SAT-V and the GRE CATs, overlap constraints, were purposely dealt with or met first, before content and psychometric constraints, to insure that rules on blocking and overlap would be violated only a small percentage of the time, if at all.

5. The results of the simulation were evaluated in a number of ways, both conditional on score level and unconditionally. For the evaluations that were conditional on score level, the same score points on the observed score metric that were used in the simulations were also used in these evaluations. At each score point, the proportion of constraint violations and the average number of items administered were calculated for each constraint. In addition, the conditional standard errors of measurement (CSEMs) were examined. The CSEMs were compared to the CSEMs for the reference test as well as the CSEMs for the reference test scaled to the number of items used for the CAT in the particular simulation.

6. For the unconditional evaluation, the item parameters and item responses from a large group of examinees (> 5000) who took the reference test were used to compute an estimated distribution of true ability using the method developed by Mislevy (1984). Proportional values of this distribution were applied to the conditional results to yield an estimate of the unconditional results in a typical group of test takers. Estimated reliability was then computed using the method of Green et al. (1984, equation 6). In addition, a weighted total proportion of constraint violations and an overall average number of items administered were calculated for each constraint.

7. The conditional and unconditional results having to do with proportion of constraint violations, average number of items administered, and CSEMs and the

overall reliability estimate were shared with test development specialists. If the results were found to be unacceptable, weights were altered on those constraints having the largest proportion of constraint violations and test length was increased (at least in the early iterations). In addition, upper and lower bounds for a particular constraint may have been altered (again, in the early iterations).

8. The entire simulation process was repeated using the new weights, and possibly, the new test length and new lower and upper bounds on particular constraints. The results were again shared with test development specialists and if anything was found to be unacceptable, additional adjustments were made, and another iteration of the simulation process was performed.

9. When a simulation was finally performed for which the proportion of constraint violations, average number of items administered for each constraint, and CSEMs and overall estimated reliability were found to be acceptable, one additional step was performed as a final evaluation of the adaptive test design. A number of the simulated adaptive tests were printed in paper-and-pencil form for examination by test specialists not involved in CAT development. All reviews were performed blind, that is, the specialists had no knowledge of the content constraint violations, the specifications for overlap, or the ability levels for which the CATs were appropriate.

10. If the set of tests passed test specialist review and a separate test sensitivity review, the iterative process was stopped and the pool of items taken into the final simulation was designated as the pool to be used when the CAT was to be administered operationally.

RESULTS

The results that follow are based on the final iteration of the sequence of simulations that were performed during the design phase of each of the five CATs described in this paper. Again, SAT results are described prior to GRE General results because of the sequential way in which the five CATs were developed. Four categories of results will be described for the final simulated CATs: 1) psychometric properties; 2) satisfaction of content constraints; 3) exposure rates of items; and 4) satisfaction of test development specialist reviews.

Psychometric Properties

The estimated reliabilities of all five CATs were computed using the method suggested by Green et al. (1984). In the case of the SAT CATs, the simulations were performed in order that this reliability estimate match a target reliability estimate generated by averaging the reliability estimates for the ten most recently administered SAT forms. In the case of the GRE General CATs, the simulations were performed in order that this

reliability estimate match the target reliability estimate from the full-length paper-and-pencil reference test. (For the SAT CATs, the reliability of the reference test was somewhat higher than that of the average paper-and-pencil test. The form used as the reference test for both SAT CATs was specifically chosen so that the comparability study documented in Eignor (1993) could be accomplished.)

Table 13 contains the target reliability, the estimated CAT reliability, and the estimated reference test reliability for each of the five CATs. As can be seen in Table 13, for each of the five CATs, the estimated CAT reliability reached or exceeded the target reliability.

Insert Table 13 about here

Figures 1-5 display more detail about each of the five CATs and their corresponding reference tests. Each of the figures contains information on a single CAT, and three curves are displayed in each figure: 1) the conditional standard error of measurement (CSEM) curve for the full length paper-and-pencil reference test (see Lord, 1980, equation 4.8); 2) the CSEM curve for the reference test scaled to a length of the adaptive test; and 3) the CSEM curve for the adaptive test.

Insert Figures 1-5 about here

As can be seen from the plots in Figures 1-5, the CAT in each instance has considerably smaller CSEMs than the reference test scaled to a length of the CAT, and at some score points, the CAT CSEMs approach the size of the CSEMs for the full length paper-and-pencil reference test. At other score points, particularly for the GRE CATs, the CAT CSEMs are smaller than the CSEMs for the full length reference test.

Satisfaction of Content Constraints

SAT-V

The 27-item SAT-V CAT achieved the level of precision of measurement just described without violating item overlap constraints, set constraints, or major content constraints. That is, each of the CATs simulated in the final simulation run had no overlap constraint violations and each contained exactly 8 reading comprehension items based on 2 long reading comprehension passages and one medium passage, 5 sentence completion items, 6 analogy items, and 8 antonym items, all blocked in the appropriate fashion.

While the major content constraints were not violated, some other content constraints were violated. Table 14 displays, for each constraint that had some violation, the proportion of examinees in a typical population that could be expected to experience such violations and the typical extent of such violations. The number of items administered for each constraint, averaged over the typical distribution, rarely violates the constraint. However, the conditional number of items at each ability level (not displayed in the table) shows that constraint violations tend to accrue when there is a relationship between an item with a particular feature and the appropriateness of the item for a particular ability level. For example, 28.8% of the typical population have adaptive tests that violated constraint 26 having to do with analogy items, which called for either 1 or 2 of these items to be administered. A substantial proportion of simulees with below average true ability were administered three or four of these items.

Insert Table 14 about here

The constraint violations exhibited in Table 14 could be reduced if it were possible to obtain items appropriate for all levels of ability that also had all of the features of interest. Test development staff working on the SAT-V CAT felt that the constraint violations displayed in Table 14 were sufficiently minor that there was no need to go to the additional effort of augmenting the pool.

SAT-M

The 20-item SAT-M CAT achieved the level of measurement precision just described without violating set or major content constraints. That is, each of the CATs simulated in the final simulation run had the 7 quantitative comparison items blocked together appropriately and each simulated CAT contained 7 quantitative comparison items and 13 regular five-choice problem solving items.

While major constraints were not violated, some other constraints were violated. Table 15 displays, for each constraint that had some violation, the proportion of examinees in a typical population that could be expected to experience such violations and the typical extent of such violations. Using information contained in the table and other information contained elsewhere, it may be surmised that a typical examinee will receive a total of 13 regular five-choice problem solving items, of which 3 or 4 will be arithmetic items, 4 will be algebra items, 4 will be geometry items, and 1 or 2 will be miscellaneous items. (All these constraints were met with no violations.) Further, this examinee will receive a total of 7 quantitative comparison (QC) items, but depending on ability level, may not receive the prespecified 2 QC arithmetic items or 1 QC miscellaneous item. (These last two constraints had violations.)

Insert Table 15 about here

Test development staff working on the SAT-M CAT felt that the constraint violations involving the QC items, and the other constraint violations displayed in Table 15, were sufficiently minor that there was no need for additional work to augment the pool.

GRE Verbal

The 30-item GRE Verbal CAT achieved the level of measurement precision just described without violating item overlap, passage, or major item type constraints. That is, each of the CATs simulated in the final simulation run had 3 reading comprehension passages, one long and two short, 8 reading comprehension items based on the 3 passages, 6 sentence completion items, 7 analogy items, and 9 antonym items.

While major constraints were not violated, some other constraints, some with fairly large weights, were violated. Table 16 displays, for each constraint that had some violation, the proportion of examinees in a typical population that could be expected to experience such violations and the typical extent of such violations. Most of the constraints with fairly large weights that were violated had to do with reading comprehension passages. It would appear that the total number of 31 reading comprehension passages in the pool is not sufficient to satisfy all of the constraints. Test development staff working on the GRE Verbal CAT felt, however, that these constraint violations and the others listed in Table 16 were infrequent enough and sufficiently minor that there was no need to go through the effort of augmenting the pool with additional passages and items.

Insert Table 16 about here

GRE Quantitative

The 28-item GRE Quantitative CAT achieved the level of measurement precision just described without violating item overlap, set, or major item type and content constraints. That is, each of the simulated CATs in the final simulation run had two data interpretation sets, 4 data interpretation items based on the two sets, 14 quantitative comparison (QC) items, and 10 problem solving (PS) items. In addition, all simulated CATs contained 13 arithmetic items, 8 algebra items, and 7 geometry items.

However, as with the other tests, some other more minor constraints were violated. Table 17 displays, for each constraint that had some violation, the proportion of examinees in a typical population that could be expected to experience such violations and the typical extent of such violations. All but one of the constraints that have violations have to do with

subdivisions of QC Type 1 or PS Type 1 items, and all had weights of one. Test development staff felt these constraint violations were sufficiently minor that the Quantitative pool did not need to be augmented.

Insert Table 17 about here

GRE Analytical

The 35-item GRE Analytical CAT achieved the level of measurement precision just described without violating any overlap or major set and content constraints. That is, each of the CATs simulated in the final simulation run had 6 analytical reasoning sets, a total of 26 analytical reasoning items based on these sets, and 9 logical reasoning items.

Of the 39 constraints for GRE Analytical, only two had any significant level of constraint violations. This is partly a function of the size of the final Analytical pool, which was considerably larger than the pools for the other CATs, and a function of the weights associated with the Analytical constraints. Table 18 displays data comparable to that displayed in the other tables of the same kind. Both constraints that have violations deal with specific features of the logical reasoning items. Test development staff felt these constraint violations were minor and that additional work on the pool was not needed.

Insert Table 18 about here

Item Exposure Rates

SAT CATs

Item exposure was controlled with the SAT-V and SAT-M CATs through the use of a count-down randomization procedure. The first item to be administered was randomly chosen from a list of the eight best items, the second item was randomly chosen from a list of the seven best items, the third from a list of the six best items, and so forth. The eighth and subsequent items were chosen to be optimal.

Table 19 presents the expected average exposure rates for the SAT-V CAT items and passages in the final pool when given to a typical group of test takers. In this typical administration, 272 of the items and passages in the 330 "item" pool (303 test items and 27 passages) would have been administered. The highest exposure rates for items or passages are in the .5 to .6 range, i.e., the item or passage would show up on from 50 to 60% of the CATs administered to the typical population. The average exposure rate for all used items and passages is just over 11%.

Insert Table 19 about here

Table 20 presents the expected average exposure rates for the SAT-M CAT items and sets in the final pool when given to a typical group of examinees. In this typical administration, 206 of the items and sets in the 241 "item" pool (235 test items and 6 sets) would have been administered. As with the SAT-V CAT, the highest exposure rates are in the 50 to 60 percent range and the average exposure rate for all used items and sets is just over 10%.

Insert Table 20 about here

The SAT-V and SAT-M CATs were developed before the Extended Sympson/Hetter methodology was added to the simulation system. While the exposure rates shown in Tables 19 and 20 are higher than what might have been preferred, they were deemed acceptable given the plans for how the SAT CAT is to be used. Hence, the simulations done for the SAT CAT were not redone when the Extended Sympson/Hetter methodology was implemented.

GRE General CATs

Item and passage or set exposure was controlled for each of the GRE CATs through use of the Extended Sympson/Hetter (ESH) methodology with the desired maximum rate of usage set at .2.

Table 21 presents the expected average exposure rates for the GRE Verbal items and passages in the final pool when given to a typical group of test takers. In this typical administration, 310 of the items and passages in the 381 "item" pool (350 test items and 31 passages) would have been administered. The highest exposure rates for items and passages are in the .2 to .3 range; i.e., the item or passage would have shown up on from 20 to 30% of the CATs administered to the typical population. The average exposure rate for all used items and passages was just over 10%.

Insert Table 21 about here

Figure 6 shows the maximum observed exposure rates over each of the eight iterations of the ESH procedure that led to final exposure control parameters for the ESH method that were used to generate the exposure rate data in Table 21. As can be seen in Figure 6, the exposure control parameters and, hence, the exposure rates for the discrete

items, stimuli (reading comprehension passages), and items in sets (reading comprehension items) stabilized at the fourth iteration (i.e., fourth sequential simulation).

Insert Figure 6 about here

Table 22 presents the expected average exposure rates for the GRE Quantitative items and sets in the final pool when given to a typical group of examinees. In this typical administration, 285 sets and items in the 348 "item" pool (330 test items and 18 sets) would have been administered. As with the GRE Verbal CAT, the highest exposure rates for items and sets are in the .2 to .3 range. The average exposure rate for all used items and sets was just over 10%.

Insert Table 22 about here

Figure 7 shows the maximum observed exposure rates over each of the eight iterations of the ESH procedure that led to final exposure control parameters for the ESH method that were used to generate the exposure rate data in Table 22. As can be seen in Figure 7, the exposure control parameters and, hence, the exposure rates for the discrete items, stimuli (data interpretation sets), and items in sets (data interpretation items) stabilized at the third iteration.

Insert Figure 7 about here

Table 23 presents the expected average exposure rate for the GRE Analytical items and sets in the final pool when given to a typical group of examinees. In this typical administration, 464 sets and items in the 512 "item" pool (449 test items and 63 sets) would have been administered. The highest exposure rates for items and sets are again in the .2 to .3 range. However, for GRE Analytical, the average exposure rate for all used items and sets is just under 9%.

Insert Table 23 about here

Figure 8 shows the maximum observed exposure rates over each of the eight iterations of the ESH procedure that led to final exposure control parameters for the ESH method that were used to generate the exposure rate data in Table 23. As can be seen in Figure 8, the exposure control parameters and, hence, the exposure rates for the discrete

logical reasoning items, stimuli (analytical and logical reasoning sets), and items in sets (analytical and logical reasoning items based on the sets) were quite stable through the complete sequence of eight iterations.

Insert Figure 8 about here

Test Specialist Review

As a final evaluation of the adaptive test design for the five CATs discussed in this paper, paper-and-pencil copies of actual adaptive tests were examined by test development specialists. The test reviews were performed blind, that is, the test specialists who performed the reviews had no knowledge of the content constraint violations, the specifications for overlap, or the ability levels for which the adaptive tests were appropriate.

The manner in which the adaptive tests were chosen was similar for all five tests. For example, SAT-V paper-and-pencil copies of 30 adaptive tests were examined by test development specialists. Ten of these tests were drawn randomly from those administered to simulees at the four lowest and six highest ability levels. Twelve of them were drawn randomly from simulees at the five middle ability levels (true scores of 35, 40, 45, 50, and 55) within which about 67% of the typical distribution of abilities lies. The remaining eight tests were drawn randomly from simulees who had particular patterns of content constraint violations.

A number of problems with these sample CATs for each test were identified, particularly for CATs appropriate for the more extreme ability levels as opposed to those CATs appropriate for more typical examinees. This is not surprising given the fact that many items in the pools were designed to measure best at middle ability levels; thus the pools are richest in items appropriate for these abilities. All problems were carefully investigated, and none of them could be attributed to the adaptive testing methodology employed. Rather, all problems were identified as stemming from the size, nature and characteristics of the item pools and the specifications for overlap.

OTHER CAT ISSUES

Timing

Before the SAT and GRE General CATs could be given operationally, a number of additional decisions had to be made, one of which involved timing of the CATs. The CATs should be given under unspeeded conditions, so that essentially all examinees are afforded the opportunity to complete the CATs.

After consultation with test development specialists who had estimates of the amount of time needed to complete items in the various Verbal and Math item or content categories when the items are presented in paper-and-pencil format, initial decisions were made on the timing of the SAT CATs. The 27-item SAT-V CAT had a time limit of 40 minutes while the 20-item SAT-M CAT had a time limit of 30 minutes. However, after some small scale initial pilot testing, it was found that above average examinees were having difficulty completing the 20 difficult Math items administered to them in 30 minutes. Hence, the time limit for the SAT-M CAT was increased to 40 minutes. When the CATs are administered in high schools, the time limits are to be 40 minutes for each CAT.

Unlike the SAT CATs, where timing decisions were made without the benefit of a formal study or analysis, a study was conducted for the GRE CATs to help establish the time limits (see Reese, 1993). Examinee CAT times were modeled using existing timing information from a linear computer-based GRE General form given in a recent field test. Given the results of this study and other information, recommended testing times for the GRE General CATs are as follows: Verbal-30 minutes, Quantitative-45 minutes, and Analytical-60 minutes.

Review of Responses to Items

The general delivery system in place for computerized testing at Educational Testing Service allows the possibility for examinees to go back and review their responses to items on a computerized test. This facility was essentially put in place for linear computer-based tests which are to be administered in as parallel a fashion to the paper-and-pencil test as possible. Since review (within a section) is possible for most paper-and-pencil tests, it was seen as important that review also be possible for these tests. However, the use of the review function with a CAT could cause problems for the sequential updating of ability estimates done after administration of items so subsequent items can be chosen. Review and change of a response to an item previously administered might cause all subsequent items actually administered to no longer be strictly appropriate or optimal. Research is needed on this topic, particularly for CATs that are the length of those described in this paper. Given the present lack of research, a decision was made with the SAT and GRE CATs that review not be allowed. Examinees are allowed to progress only in a forward fashion. In addition, examinees are not given the option to omit items on the CATs.

DISCUSSION

SAT and GRE General paper-and-pencil test forms, and forms for tests from other large scale testing programs, are constructed from very detailed sets of test specifications. When a program like GRE wants to introduce computer adaptive testing as an alternative to paper-and-pencil testing, it is essential that the detailed set of specifications in place for the construction of the paper-and-pencil form be applied in the construction of the computer adaptive tests. This is for two reasons: 1) so that the content tested on the CATs is similar to that tested via paper-and-pencil, thereby ensuring that an examinee who takes a CAT is in

no way disadvantaged with respect to the content of the items received; and 2) because scores from the paper-and-pencil forms and the CATs will need to be used interchangeably, it is essential that a comparability study be done. Models in place for establishing this sort of comparability of scores are based on the assumption that the scores to be equated originate from parallel measures.

Previous methodologies used for the construction of adaptive tests at Educational Testing Service are unable to take into account the number and complexity of constraints on item selection that govern paper-and-pencil test construction practice for tests like the GRE General and the SAT. Hence, until very recently, CATs that are similar to paper-and-pencil forms could not be constructed in these Programs. With the development of the Stocking/Swanson weighted deviations methodology, GRE and SAT CAT test construction has become a reality.

The success of the Stocking/Swanson methodology rests on the fact that it can incorporate content, overlap, and set constraints in the sequential selection of items as desired properties of the resultant adaptive tests, rather than as strict requirements. At the same time, the new methodology minimizes aggregate failures in the same fashion as in the construction of paper-and-pencil tests. The extent to which restrictions in item selection are not satisfied is then the result of deficiencies in the item pool, as it is with the paper-and-pencil test.

When the new methodology for constructing CATs is coupled with simulation procedures, a number of distinct advantages accrue. As can be seen from the description in this paper of the development of the SAT and GRE CATs, simulation allows the developer to pinpoint many of the necessary characteristics of the CAT prior to actual administration to real examinees. In addition, the simulation results provide very reasonable expectations for how the CATs will work when given to these real examinees. Output from the simulations can be used to provide initial psychometric documentation for the CAT. Finally, simulation results can provide needed data for further research and analyses. For example, simulated examinee response strings may be used as data to quality control an entire CAT delivery system prior to administration with real examinees. In short, it would be difficult to argue with the assertion that the CAT development process requires the use of in-depth simulation procedures.

While on all counts it would appear that use of the Stocking/Swanson methodology coupled with simulation procedures has led to viable SAT and GRE General CATs, the proof will come in actual administration of the CATs to examinees. Currently, the SAT CAT is being administered at a number of high schools. The GRE General CAT is just beginning to be administered as part of the variable section of the linear computerized test developed and administered in that Program. Feedback from examinees and test users will ultimately determine the success of the two CAT development efforts.

REFERENCES

Ackerman, T. (1989, March). An alternative methodology for creating parallel test forms using the IRT information function. Paper presented at the annual meeting of NCME, San Francisco.

Eignor, D. R. (1993, April). Deriving comparable scores for computer adaptive and conventional tests: An example using the SAT. Paper presented at the annual meeting of NCME, Atlanta.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 21, 347-360.

Henrysson, S. (1971). Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, DC: American Council on Education.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. Applied Psychological Measurement, 1, 95-100.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), New horizons in testing. New York: Academic Press.

Mislevy, R. J. (1984). Estimating latent distributions. Psychometrika, 49, 359-381.

Reese, C. M. (1993, April). Establishing time limits for the GRE computer adaptive tests. Paper presented at the annual meeting of NCME, Atlanta.

Stocking, M. L. (1987). Two simulated feasibility studies in computerized adaptive testing. Applied Psychology: An International Review, 36, 3/4, 263-277.

Stocking, M. L. (1993). Controlling item exposure rates in a realistic adaptive testing paradigm (RR-93-2). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Stocking, M. L., & Swanson, L. (1992). A method for severely constrained item selection in adaptive testing (RR-92-37). Princeton, NJ: Educational Testing Service.

Swanson, L., & Stocking, M. L. (1992). A model and heuristic for solving very large item selection problems (RR-92-31). Princeton, NJ: Educational Testing Service.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. Proceedings of the 27th annual meeting of the Military Testing Association. San Diego, CA: Navy Personnel Research and Development Center.

TD/DC 5.0 User's Manual (1990). Copyright 1989, 1990, 1991. Princeton, NJ: Educational Testing Service.

Theunissen, T. J. J. M. (1985). Binary programming and test design. Psychometrika, 50, 411-420.

Theunissen, T. J. J. M. (1986). Some applications of optimization algorithms in test design and adaptive testing. Applied Psychological Measurement, 10, 381-389.

van der Linden, W. J. (1987). Automated test construction using minimax programming. In W. J. van der Linden (Ed.), IRT-based test construction. Enschede, The Netherlands: Department of Education, University of Twente.

van der Linden, W. J., & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. Psychometrika, 54, 237-248.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia.

Table 1

Comparison of Full Length SAT-Verbal Test

and SAT-Verbal Computer Adaptive Test

RC: Reading Comprehension Items

ANT: Antonym Items

ANAL: Analogy Items

SC: Sentence Completion Items

Numbers of Items

	RC	ANT	ANAL	SC	TOTAL
PAPER AND PENCIL	25[1]	25	20	15	85
CAT	8[2]	8	6	5	27
VERBAL POOL	91[3]	74	51	87	303

[1]Based on 5 or 6 passages with 3 to 5 items per passage

[2]Based on 3 passages; two passages have 3 items each, one passage has two items

[3]Based on 27 passages, having from 3 to 6 items per passage

Table 2

Comparison of Full Length SAT-Math Test

and SAT-Math Computer Adaptive Test

PS: Regular 5-Choice Problem Solving Math Items

QC: 4-Choice Quantitative Comparison Items

ARIT: Arithmetic Items

ALGB: Algebra Items

GEOM: Geometry Items

MISC: Miscellaneous Items

Numbers of Items

	PS	QC	ARIT	ALGB	GEOM	MISC	TOTAL
PAPER AND PENCIL	40	20	18-19	17	16-17	7-9	60
CAT	13	7	5-6	6	6	2-3	20
MATH POOL	128	107	70	65	66	34	235

BEST COPY AVAILABLE

Table 3

Comparison of Full Length GRE General Verbal Test

and GRE Verbal Computer Adaptive Test

RC: Reading Comprehension Items

ANT: Antonym Items

ANAL: Analogy Items

SC: Sentence Completion Items

Numbers of Items

	RC	ANT	ANAL	SC	TOTAL
PAPER AND PENCIL	22[1]	22	18	14	76
CAT	8[2]	9	7	6	30
VERBAL POOL	185[3]	68	52	45	350

[1]Based on 4 passages, having from 4 to 7 items per passage

[2]Based on 3 passages, the two short passages having 2 items each, the one long passage having 4 items

[3]Based on 31 passages, having from 5 to 10 items per passage

BEST COPY AVAILABLE

Table 4

Comparison of Full Length GRE General Quantitative Test

and GRE Quantitative Computer Adaptive Test

PS: 5-Choice Problem Solving Items

QC: 4-Choice Quantitative Comparison Items

DI: 5-Choice Data Interpretation Items

ARIT: Arithmetic Items

ALGB: Algebra Items

GEOM: Geometry Items

Numbers of Items

	PS	QC	DI	ARIT	ALGB	GEOM	TOTAL
PAPER AND PENCIL	20	30	10[1]	22-32	13-21	12-20	60
CAT	10	14	4[2]	13	8	7	28
QUANTITATIVE POOL	81	120	129[3]	200	70	60	330

[1]Based on 2 sets, each set having 5 items

[2]Based on 2 sets, each set having 2 items

[3]Based on 18 sets, having from 5 to 11 items per set

Table 5

Comparison of Full Length GRE General Analytical Test

and GRE Analytical Computer Adaptive Test

AR: Analytical Reasoning Items

LR: Logical Reasoning Items

Numbers of Items

	AR	LR	Total
PAPER AND PENCIL	38[1]	12	50
CAT	26[2]	9	35
ANALYTICAL POOL	374[3]	75[4]	449

[1]Based on no more then 6 sets, each set having from 3 to 8 items each

[2]Based on 6 sets, having from 4 to 5 items per set

[3]Based on 61 sets, having from 6 to 8 items per set

[4]Based on 69 discrete items and 3 sets of 2 items each

Table 6

Content Constraints and Weights for the SAT-Verbal Computer Adaptive Test

Number	Description	LB[1]	UB[2]	W[3]	n[4]	Number	Description	LB[1]	UB[2]	W[3]	n[4]
1	Long Reading Comprehension Passages (RCP)	2	2	20	17	22	SC-2	3	3	20	53
2	Medium RCP	1	1	20	10	23	SC-a	0	1	20	12
3	RCP-A[5]	1	1	20	8	24	SC-b	2	2	20	22
4	RCP-B	0	1	2	6	25	Analogy (ANAL) Items	6	6	20	51
5	RCP-C	0	1	1	3	26	ANAL-A	1	2	1	12
6	RCP-1	0	1	1	6	27	ANAL-B	1	2	1	12
7	RCP-2	0	1	1	4	28	ANAL-C	1	2	1	13
8	RCP-a	0	1	20	6	29	ANAL-D	1	2	1	14
9	RCP-b	0	1	1	1	30	ANAL-1	1	3	1	12
10	RCP-I	1	1	20	7	31	ANAL-2	1	3	1	20
11	Reading Comprehension (RC) Items	8	8	20	91	32	ANAL-a	0	1	1	8
12	RC-A	1	4	1	13	33	ANAL-I	0	1	1	5
13	RC-B	1	4	1	27	34	Antonym (ANT) Items	8	8	20	74
14	RC-C	2	5	1	25	35	ANT-A	1	2	1	13
15	RC-D	1	4	1	26	36	ANT-B	1	2	3	20
16	Sentence Completion (SC) Items	5	5	20	87	37	ANT-C	1	2	3	19
17	SC-A	1	2	3	23	38	ANT-D	1	2	1	22
18	SC-B	1	2	3	24	39	ANT-1	1	4	1	24
19	SC-C	1	2	3	21	40	ANT-2	1	4	1	15
20	SC-D	1	2	3	19	41	ANT-3	1	4	1	35
21	SC-1	2	2	20	34						

[1]LB = Lower Bound; [2]UB = Upper Bound; [3]W = Weight; [4]n = Number in Pool

[5]Naming Conventions: Capital letters (or numbers, small letters, and roman numerals) indicate sub-strata of items or passages that collectively may or may not exhaust the strata. Sets of items indicated by, say, letters are not mutually exclusive from sets of items indicated by, say, numbers, i.e., an item may satisfy more than one constraint.

BEST COPY AVAILABLE

Table 7

A Portion of the Overlap Groups for the SAT-Verbal Computer Adaptive Test

Group Number	Number in Group	Items/Passages in Group
1	4	232, 22, 242, 103
2	3	232, 218, 79
3	3	232, 298, 307
.	.	.
.	.	.
250	3	321, 284, 281
251	4	321, 305, 281, 308
252	3	38, 240, 142
.	.	.
.	.	.
526	2	449, 550
527	2	518, 556
528	2	518, 565

Table 8

A Portion of the List of Blocks for the SAT-Verbal Computer Adaptive Test

Block	Number to Select	Starting Position	Ending Position	Classification
1	5	1	95	SC
2	6	96	180	ANAL
3	8	181	321	ANT
4	3	322	327	Long RCP
5	3	328	333	Long RCP
.	.	.	.	.
.	.	.	.	.
52	2	556	559	Medium RCP
53	2	560	564	Medium RCP
54	3	565	569	Medium RCP

Table 9

Content Constraints and Weights for the SAT-Math Computer Adaptive Test

Number	Description	LB[1]	UB[2]	W[3]	n[4]	Number	Description	LB[1]	UB[2]	W[3]	n[4]
1	Item Sets	0	1	20	6	28	PS ARIT-B	1	1	11	13
2	Items from all content areas (ALL)-A[5]	4	5	20	57	29	PS ARIT-B1[6]	0	1	1	1
3	ALL-B	15	16	20	178	30	PS ARIT-B2	0	1	1	4
4	ALL-1	0	1	2	8	31	PS ARIT-B3	0	1	1	1
5	ALL-2	0	1	2	9	32	PS ARIT-B4	0	1	1	5
6	ALL-3	0	2	2	11	33	PS ARIT-B5	0	1	1	2
7	ALL-4	0	2	2	4	34	PS ARIT-C	0	1	10	6
8	ALL-5	0	1	2	3	35	PS ARIT-C1	0	1	1	4
9	ALL-6	0	1	1	8	36	PS ARIT-C2	0	1	1	2
10	4-Choice and 5-Choice Geometry Items (ALL GEOM)-A	1	3	1	16	37	PS ARIT-D	0	1	10	3
11	ALL GEOM-B	0	3	1	12	38	PS ARIT-E	0	1	10	3
12	ALL GEOM-C	2	6	1	37	39	PS ARIT-F	0	1	10	1
13	ALL GEOM-D	0	2	1	14	40	PS ARIT-G	0	1	10	2
14	ALL GEOM-E	0	1	1	3	41	PS ARIT-H	1	2	20	9
15	ALL GEOM-F	0	1	2	5	42	PS-Algebra (ALGB) Items	4	4	20	34
16	4-Choice Quantitative Comparison (QC) Items	7	7	20	107	43	PS ALGB-A	0	1	10	1
17	QC-Arithmetic Items	2	2	20	32	44	PS ALGB-B	2	2	11	17
18	QC-Algebra Items	2	2	20	31	45	PS ALGB-B1	0	1	1	1
19	QC-Geometry Items	2	2	20	28	46	PS ALGB-B2	0	1	1	1
20	QC-Miscellaneous Items	1	1	20	16	47	PS ALGB-C	0	1	10	2
21	QC-A	1	2	1	27	48	PS ALGB-D	0	1	10	2
22	QC-B	1	2	1	18	49	PS ALGB-E	1	1	11	7
23	QC-C	1	2	2	29	50	PS ALGB-F	0	1	10	2
24	QC-D	1	2	2	33	51	PS ALGB-G	0	1	10	2
25	5-Choice Problem Solving (PS) Items	13	13	20	128	52	PS ALGB-H	0	1	10	1
26	PS-Arithmetic (ARIT) Items	3	4	20	38	53	PS-Geometry (GEOM) Items	4	4	20	38
27	PS ARIT-A	0	1	10	1	54	PS GEOM-A	0	1	10	5

[1]LB = Lower Bound; [2]UB = Upper Bound; [3]W = Weight; [4]n = Number in Pool

[5]Naming Conventions: Capital letters (or numbers) indicate sub-strata of items that collectively may or may not exhaust the strata. Sets of items indicated by letters are not mutually exclusive from sets of items indicated by numbers, i.e., an item may satisfy more than one constraint.

[6]A number following a letter, like B1, indicates that the items satisfying the constraint are a subset of the larger set of items satisfying the related constraint, i.e., the set of items satisfying constraint B1 are a subset of the items satisfying constraint B.

BEST COPY AVAILABLE

Table 9 (Con't)

Content Constraints and Weights for the SAT-Math Computer Adaptive Test

Number	Description	LB[1]	UB[2]	W[3]	n[4]	Number	Description	LB[1]	UB[2]	W[3]	n[4]
55	PS GEOM-B	0	1	10	6	66	PS GEOM-F2	0	1	1	3
56	PS GEOM-B1	0	1	1	3	67	PS GEOM-F3	0	1	1	1
57	PS GEOM-B2	0	1	1	1	68	PS GEOM-G	0	1	10	1
58	PS GEOM-B3	0	1	1	2	69	PS GEOM-H	0	1	10	6
59	PS GEOM-C	0	1	10	1	70	PS GEOM-I	0	1	10	4
60	PS GEOM-C1	0	1	1	1	71	PS Miscellaneous (MISC) Items	1	2	20	18
61	PS GEOM-D	0	1	10	4	72	PS MISC-A	0	1	1	1
62	PS GEOM-D1	0	1	1	4	73	PS MISC-B	0	1	1	1
63	PS GEOM-E	0	1	10	5	74	PS MISC-C	0	2	20	6
64	PS GEOM-F	0	1	10	6	75	PS MISC-D	0	2	20	10
65	PS GEOM-F1	0	1	1	2						

[1]LB = Lower Bound; [2]UB = Upper Bound; [3]W = Weight; [4]n = Number in Pool

Table 10

Content Constraints and Weights for the GRE General Verbal Computer Adaptive Test

Number	Description	LB[1]	UB[2]	W[3]	n[4]
1	Long Reading Comprehension Passages (RCP)	1	1	15	11
2	Short RCP	2	2	15	20
3	RCP-A[5]	0	1	10	6
4	RCP-B	0	1	5	5
5	RCP-1	1	1	10	9
6	RCP-2	1	1	10	11
7	RCP-a	1	1	10	11
8	RCP-b	1	1	10	10
9	RCP-c	1	1	10	12
10	Reading Comprehension (RC) Items	8	8	10	185
11	RC-A	1	4	1	30
12	RC-B	1	4	1	49
13	RC-C	1	4	1	58
14	RC-D	1	4	1	36
15	Sentence Completion (SC) Items	6	6	10	45
16	SC-A	0	2	1	9
17	SC-B	0	2	1	13
18	SC-C	0	2	1	15
19	SC-D	0	2	1	8
20	Analogy (ANAL) Items	7	7	10	52
21	ANAL-A	0	2	1	8
22	ANAL-B	0	2	1	16
23	ANAL-C	0	2	1	14
24	ANAL-D	0	2	1	14
25	Antonym (ANT) Items	9	9	10	68
26	ANT-A	0	3	1	15
27	ANT-B	0	3	1	20
28	ANT-C	0	3	1	11
29	ANT-D	0	3	5	22
30	All Items (ALL)-A	0	4	10	40
31	ALL-B	0	4	5	28
32	ALL-1	2	22	1	60
33	ALL-2	2	22	1	74
34	ALL-3	2	22	1	67
35	ALL-4	2	22	1	76
36	ALL-5	2	22	1	73

[1]LB = Lower Bound; [2]UB = Upper Bound; [3]W = Weight; [4]n = Number in pool

[5]Naming Conventions: Capital letters (or numbers and small letters) indicate sub-strata of items or passages that collectively may or may not exhaust the strata. Sets of items indicated by letters are not mutually exclusive from sets of items indicated by numbers, i.e., an item may satisfy more than one constraint.

BEST COPY AVAILABLE

Table 11

Content Constraints and Weights for the GRE General

Quantitative Computer Adaptive Test

Number	Description	LB[1]	UB[2]	W[3]	n[4]
1	Data Interp (DI) Sets	2	2	11	18
2	Quantitative Comparison (QC)-Arithmetic (ARIT) Items	5	5	10	40
3	QC-Algebra (ALGB) Items	5	5	10	45
4	QC-Geometry (GEOM) Items	4	4	10	35
5	Problem Solving (PS)-ARIT Items	4	4	10	31
6	PS-ALGB Items	3	3	10	25
7	PS-GEOM Items	3	3	10	25
8	Type 1 Items	9	9	1	176
9	DI Type 1 Items	4	4	1	129
10	QC Type 1 Items	3	3	10	25
11	QC Type 1-A	0	1	1	8
12	QC Type 1-B	0	1	1	9
13	QC Type 1-C	0	1	1	8
14	PS Type 1 Items	2	2	10	22
15	PS Type 1-A	0	1	1	11
16	PS Type 1-B	0	1	1	6
17	PS Type 1-C	0	1	1	5
18	All Items (ALL)-A	0	1	1	7
19	ALL-B	0	1	1	6
20	ALL-1	1	14	1	48
21	QC-A	1	11	1	28
22	QC-B	1	11	1	32
23	QC-C	1	11	1	32
24	QC-D	1	11	1	28

[1]LB = Lower Bound; [2]UB = Upper Bound; [3]W = Weight; [4]n = Number in Pool

[5]Naming Conventions: Capital letters (or numbers) indicate sub-strata of items that collectively may or may not exhaust the strata. Sets of items indicated by letters are not mutually exclusive from sets of items indicated by numbers, i.e., an item may satisfy more than one constraint.

Table 12

Content Constraints and Weights for the GRE General Analytical Computer Adaptive Test

Number	Description	LB[1]	UB[2]	W[3]	n[4]	Number	Description	LB[1]	UB[2]	W[3]	n[4]
1	Analytical Reasoning Sets (ARS)	6	6	30	61	21	LR-c	0	2	5	11
2	ARS-A[5]	2	4	5	31	22	LR-d	0	2	5	1
3	ARS-A1[6]	1	2	10	15	23	LR-e	1	3	7	15
4	ARS-A2	1	2	5	16	24	LR-f	0	3	5	9
5	ARS-B	2	2	20	19	25	LR-g	1	3	5	19
6	ARS-C	0	1	10	6	26	LR-h	0	2	5	2
7	ARS-D	0	1	10	5	27	LR-i	0	2	5	4
8	ARS-1	0	1	5	3	28	LR-j	0	2	5	3
9	ARS-2	0	1	5	3	29	LR-I	0	3	10	14
10	ARS-a	0	1	5	2	30	LR-II	0	3	10	15
11	ARS-b	0	1	5	1	31	LR-III	0	3	10	5
12	ARS-c	0	2	10	10	32	LR-IV	0	3	5	6
13	Analytical Reasoning (AR) Items	26	26	30	374	33	All Items (ALL)-A	0	5	5	11
14	Logical Reasoning (LR) Items	9	9	30	75	34	ALL-B	0	5	5	11
15	LR-A	4	5	5	38	35	ALL-1	1	31	1	83
16	LR-B	4	5	5	37	36	ALL-2	1	31	1	95
17	LR-C	1	3	5	18	37	ALL-3	1	31	1	92
18	LR-1	0	1	5	3	38	ALL-4	1	31	1	96
19	LR-a	0	2	5	1	39	ALL-5	1	31	1	83
20	LR-b	2	3	7	10						

[1]LB = Lower Bound; [2]UB = Upper Bound; [3]W = Weight; [4]n = Number in Pool

[5]Naming Conventions: Capital letters (or numbers, small letters, and roman numerals) indicate sub-strata of items or sets that collectively may or may not exhaust the strata. Sets of items (or sets) indicated by, say, letters are not mutually exclusive from sets of items indicated by, say, numbers, i.e., an item may satisfy more than one constraint.

[6]A number following a letter, like A1, indicates that the sets satisfy the constraint are a subset of the larger set of sets satisfying the related constraint; i.e., the set of sets satisfying constraint A1 is a subset of the sets satisfying constraint A.

BEST COPY AVAILABLE

Table 13

Target CAT Reliability and Estimated CAT

and Reference Test Reliabilities

Test

Reliability	SAT-V	SAT-M	GRE General Verbal	GRE General Quantitative	GRE General Analytical
Target	.91	.91	.89	.92	.89
Estimated CAT	.91	.92	.90	.93	.89
Estimated Reference Test	.93	.93	.89	.92	.89

Table 14

Content Constraint Violations for the SAT-Verbal Computer Adaptive Test

Number	Description	LB[1]	UB[2]	W[3]	n[4]	Percent in Typical Group	Average Number of Items
4	RCP-B	0	1	2	6	3.2	.47
7	RCP-2	0	1	1	4	12.3	.77
12	RC-A	1	4	1	13	34.6	.74
13	RC-B	1	4	1	27	18.6	1.62
14	RC-C	2	5	1	25	11.2	2.46
15	RC-D	1	4	1	26	16.0	3.18
17	SC-A	1	2	3	23	12.1	1.23
18	SC-B	1	2	3	24	13.9	1.14
19	SC-C	1	2	3	21	9.3	1.46
20	SC-D	1	2	3	19	12.4	1.16
26	ANAL-A	1	2	1	12	28.8	1.86
27	ANAL-B	1	2	1	12	1.8	1.25
28	ANAL-C	1	2	1	13	4.9	1.25
29	ANAL-D	1	2	1	14	31.5	1.64
30	ANAL-1	1	3	1	12	1.5	2.19
31	ANAL-2	1	3	1	20	9.5	1.91
32	ANAL-I	0	1	1	8	3.5	.57
35	ANT-A	1	2	1	13	31.5	1.60
36	ANT-B	1	2	1	20	4.4	1.77
37	ANT-C	1	2	1	19	29.8	2.05
38	ANT-D	1	2	1	22	57.6	2.58
40	ANT-2	1	4	1	15	3.2	1.96
41	ANT-3	1	4	1	35	22.1	3.85

[1]LB = Lower Bound; [2]UB = Upper Bound; [3]W = Weight; [4]n = Number in Pool

BEST COPY AVAILABLE

Table 15

Content Constraint Violations for the SAT-Math Computer Adaptive Test

Number	Description	LB[1]	UB[2]	W[3]	n[4]	Percent in Typical Group	Average Number of Items
4	ALL-1	0	1	2	8	5.7	.53
5	ALL-2	0	1	2	9	7.5	.54
9	ALL-6	0	1	1	8	17.3	.75
10	ALL GEOM-A	1	3	1	16	25.7	1.20
13	ALL GEOM-D	0	2	1	14	12.0	1.41
15	ALL GEOM-F	0	1	2	5	1.1	.23
17	QC-Arithmetic	2	2	20	32	1.2	1.99
18	QC-Miscellaneous	1	1	20	16	1.3	1.01
21	QC-A	1	2	1	27	28.0	2.12
22	QC-B	1	2	1	18	4.7	1.53
23	QC-C	1	2	2	29	5.6	1.59
24	QC-D	1	2	2	33	5.7	1.75

[1]LB = Lower Bound; [2]UB = Upper Bound; [3]W = Weight; [4]n = Number in Pool

Table 16

Content Constraint Violations for the GRE General Verbal Computer Adaptive Test

Number	Description	LB[1]	UB[2]	W[3]	n[4]	Percent in Typical Group	Average Number of Items
5	RCP-1	1	1	10	9	3.3	1.03
6	RCP-2	1	1	10	11	2.8	.99
7	RCP-a	1	1	10	11	2.2	.98
8	RCP-b	1	1	10	10	2.0	.98
9	RCP-c	1	1	10	12	1.5	.99
11	RC-A	1	4	1	30	26.8	1.05
12	RC-B	1	4	1	49	5.7	2.36
13	RC-C	1	4	1	58	4.1	2.94
14	RC-D	1	4	1	36	26.9	1.35
16	SC-A	0	2	1	9	4.5	1.31
17	SC-B	0	2	1	13	4.2	1.49
18	SC-C	0	2	1	15	24.6	2.04
19	SC-D	0	2	1	8	4.2	1.16
21	ANAL-A	0	2	1	8	2.9	1.08
22	ANAL-B	0	2	1	16	41.5	2.42
23	ANAL-C	0	2	1	14	3.0	1.60
24	ANAL-D	0	2	1	14	20.8	1.90
26	ANT-A	0	3	1	15	5.3	1.95
27	ANT-B	0	3	1	20	18.5	2.61
29	ANT-D	0	3	5	22	14.2	2.95

[1]LB = Lower Bound; [2]UB = Upper Bound; [3]W = Weight; [4]n = Number in Pool

Table 17

Content Constraint Violations for the GRE General Quantitative Computer Adaptive Test

Number	Description	LB[1]	UB[2]	W[3]	n[4]	Percent in Typical Group	Average Number of Items
11	QC Type 1-A	0	1	1	8	28.4	1.18
12	QC Type 1-B	0	1	1	9	8.7	.82
13	QC Type 1-C	0	1	1	8	18.1	1.00
15	PS Type 1-A	0	1	1	11	2.1	.53
16	PS Type 1-B	0	1	1	6	5.6	.74
17	PS Type 1-C	0	1	1	5	3.5	.73
21	QC-A	1	11	1	28	1.9	2.81

[1]LB = Lower Bound; [2]UB = Upper Bound; [3]W = Weight; [4]n = Number in Pool

Table 18

Content Constraint Violations for the GRE General Analytical Computer Adaptive Test

Number	Description	LB[1]	UB[2]	W[3]	n[4]	Percent in Typical Group	Average Number of Items
18	LR-1	0	1	5	3	1.7	.44
20	LR-b	2	3	7	10	27.5	1.77

[1]LB = Lower Bound; [2]UB = Upper Bound; [3]W = Weight; [4]n = Number in Pool

Table 19

Item and Passage Exposure Rates for Final SAT-Verbal Pool

Exposure Rate	f	PCT.
.901 - 1.000	0	0.0
.801 - .900	0	0.0
.701 - .800	0	0.0
.601 - .700	0	0.0
.501 - .600	3	1.1
.401 - .500	9	3.3
.301 - .400	18	6.6
.201 - .300	21	7.7
.101 - .200	52	19.1
.001 - .100	169	62.1

Mean: .1103
Standard Deviation: .1230
n: 272
Number of Items and Passages: 58
Not Used

Table 20

Item and Set Exposure Rates for Final SAT-Math Pool

Exposure Rate	f	PCT.
.901 - 1.000	0	0.0
.801 - .900	0	0.0
.701 - .800	0	0.0
.601 - .700	0	0.0
.501 - .600	2	1.0
.401 - .500	3	1.5
.301 - .400	7	3.4
.201 - .300	18	8.7
.101 - .200	47	22.8
.001 - .100	129	62.6

Mean: .1013
Standard Deviation: .1036
n: 206
Number of Items and Sets: 35
Not Used

BEST COPY AVAILABLE

Table 21

Item and Passage Exposure Rates for Final GRE General Verbal Pool

Exposure Rate	f	PCT.
.901 - 1.000	0	0.0
.801 - .900	0	0.0
.701 - .800	0	0.0
.601 - .700	0	0.0
.501 - .600	0	0.0
.401 - .500	0	0.0
.301 - .400	0	0.0
.201 - .300	38	12.3
.101 - .200	116	37.4
.001 - .100	156	50.3

Mean: .1065
Standard Deviation: .0749
n: 310
Number of Items and Passages: 71
Not Used

Table 22

Item and Set Exposure Rates for Final GRE General Quantitative Pool

Exposure Rate	f	PCT.
.901 - 1.000	0	0.0
.801 - .900	0	0.0
.701 - .800	0	0.0
.601 - .700	0	0.0
.501 - .600	0	0.0
.401 - .500	0	0.0
.301 - .400	0	0.0
.201 - .300	51	17.9
.101 - .200	82	28.8
.001 - .100	152	53.3

Mean: .1053
Standard Deviation: .0763
n: 285
Number of Items and Sets: 63
Not Used

Table 23

Item and Set Exposure Rates for Final GRE General Analytical Pool

Exposure Rate	f	PCT.
.901 - 1.000	0	0.0
.801 - .900	0	0.0
.701 - .800	0	0.0
.601 - .700	0	0.0
.501 - .600	0	0.0
.401 - .500	0	0.0
.301 - .400	0	0.0
.201 - .300	22	4.7
.101 - .200	161	34.7
.001 - .100	281	60.6

Mean: .0888
Standard Deviation: .0669
n: 464
Number of Items and Sets: 48
Not Used

Figure 1

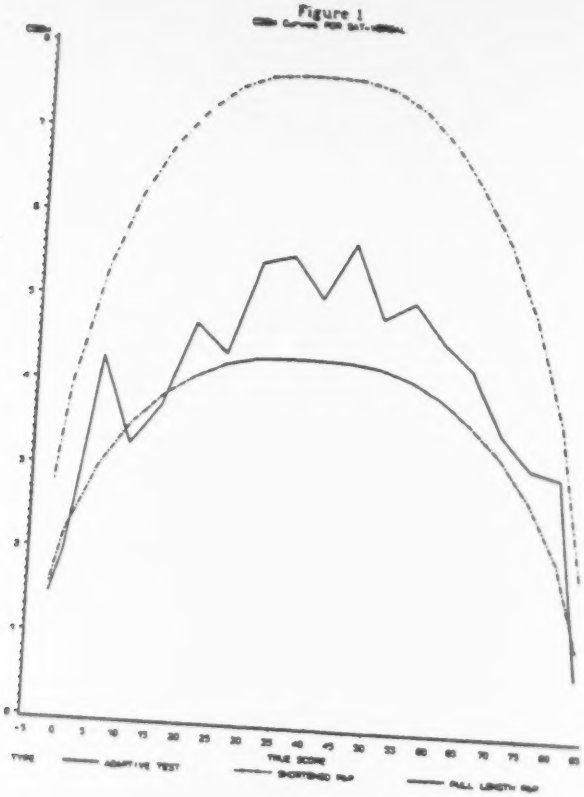


Figure 2

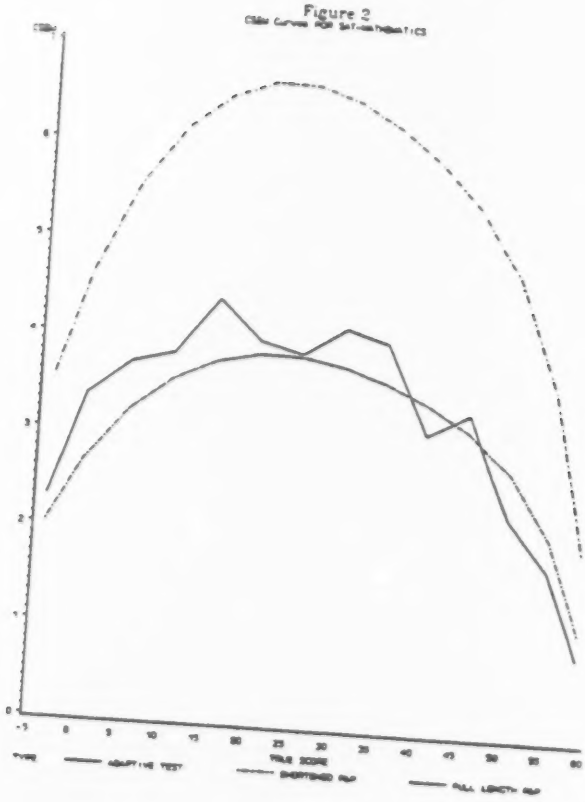


Figure 3

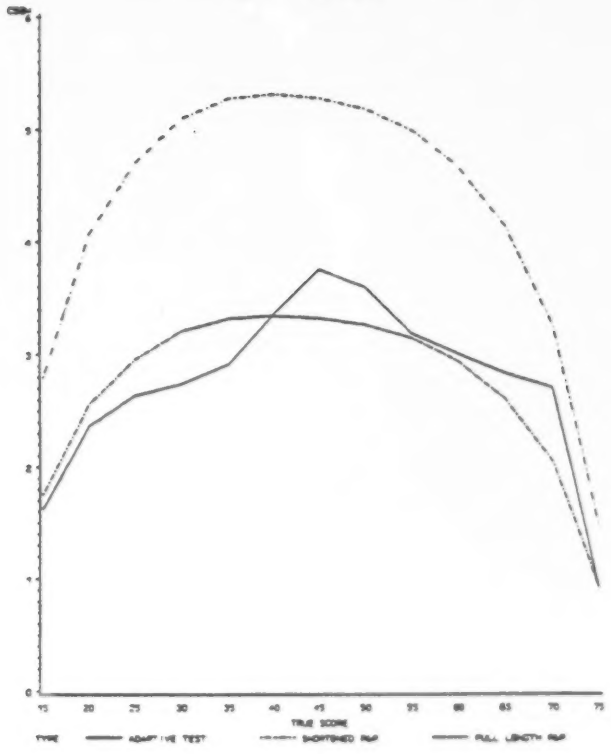


Figure 4

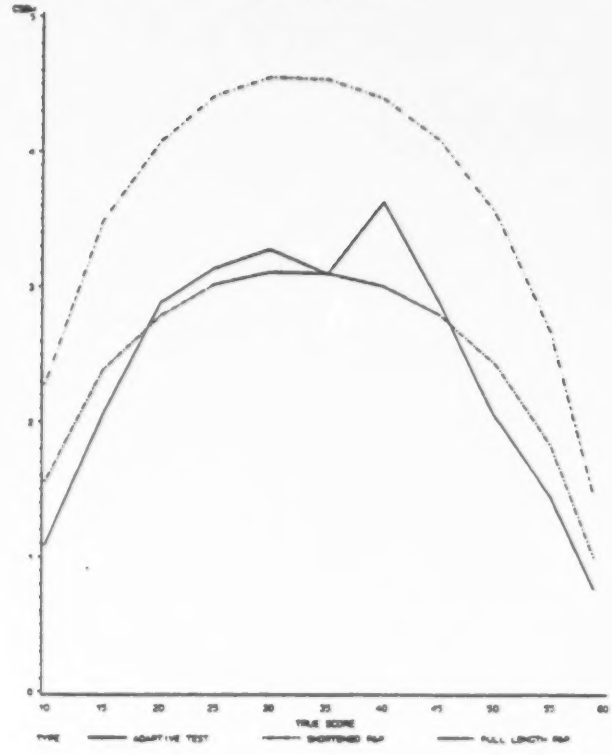
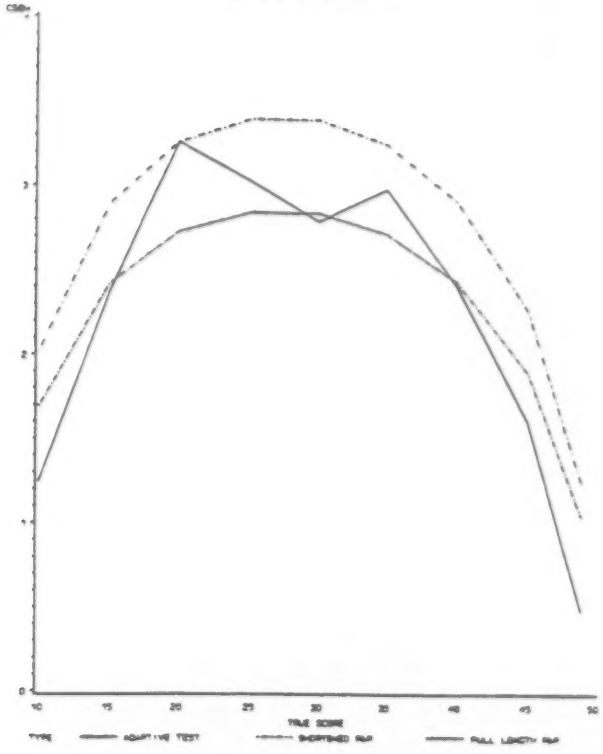


Figure 5



BEST COPY AVAILABLE

Figure 6

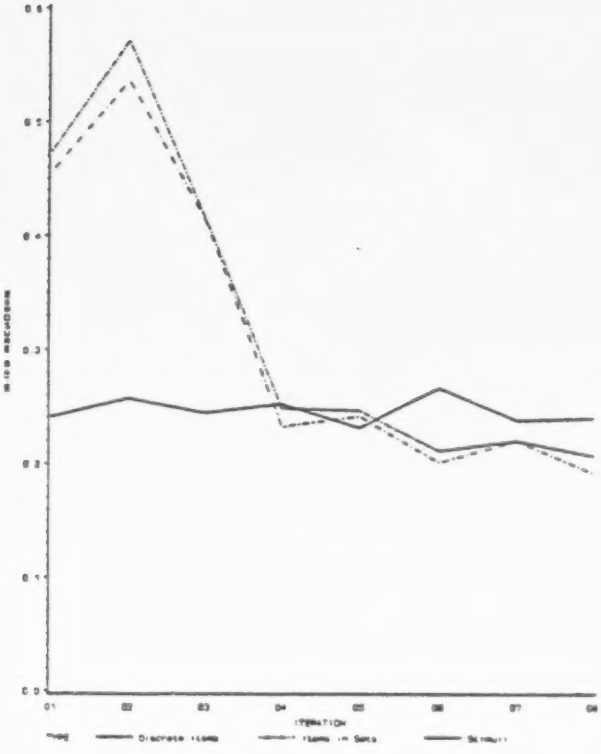


Figure 7

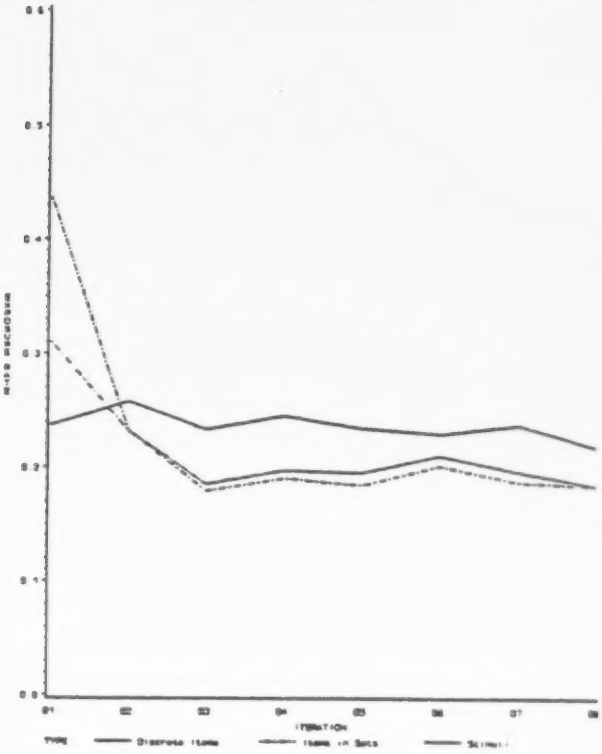
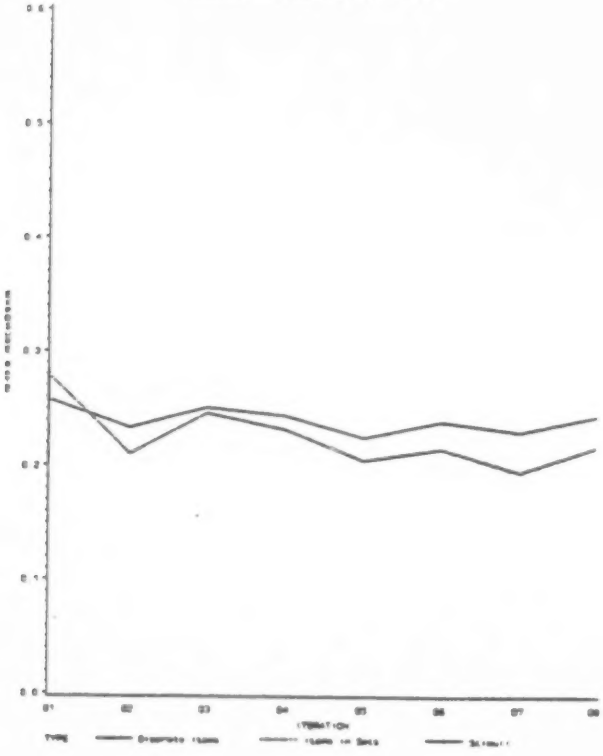


Figure 8



BEST COPY AVAILABLE